На правах рукописи

Исаченко Роман Владимирович

Снижение размерности пространства в задачах декодирования сигналов

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2021

Работа выполнена на Кафедре интеллектуальных систем Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский институт)».

Научный руководитель: **Стрижов Вадим Викторович**
доктор физико-математических наук, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, отдел интеллектуальных систем, ведущий научный сотрудник.

Официальные оппоненты: **Чуличков Алексей Иванович**
доктор физико-математических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М. В. Ломоносова», профессор кафедры математического моделирования и информатики физического факультета.

Зайцев Алексей Алексеевич
кандидат физико-математических наук, Автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий», руководитель лаборатории в Центре по научным и инженерным вычислительным технологиям для задач с большими массивами данных.

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Защита состоится 6 февраля 2020 года в 13:00 на заседании диссертационного совета Д 002.073.05 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке Федерального государственного учреждения Федеральный исследовательский центр «Информатика и управление» Российской академии наук и на сайте http://www.frccsc.ru/

Автореферат разослан декабря 2021 года.

И. о. ученого секретаря
диссертационного совета Д 002.073.05
д.т.н. И. А. Матвеев

Общая характеристика работы

Актуальность темы. В работе решается задача декодирования сигналов. Процесс декодирования заключается в восстановлении зависимости между двумя гетерогенными наборами данных. Модель предсказывает отклик на входной исходный сигнал. При построении модели возникает задача построения признакового пространства.

Исследуется проблема избыточного исходного описания данных. Исходное признаковое пространство является мультикоррелированным. При высокой мультикорреляции финальная прогностическая модель оказывается неустойчивой. Для построения простой, устойчивой модели применяются методы снижения размерности пространства [?, ?, ?] и выбора признаков [?, ?].

В работе рассматривается задача с векторной целевой переменной. Пространство целевых сигналов обладает избыточной размерностью. Методы снижения размерности, не учитывающие зависимости в целевом пространстве, являются не адекватными. При предсказании векторной целевой переменной анализируется структура целевого пространства. Предложены методы, которые учитывают зависимости как в пространстве исходных объектов, так и в пространстве целевой переменной. Предлагается отобразить пространства исходных и целевых сигналов в скрытые подпространства меньшей размерности. Для построения оптимальной модели предлагаются методы согласования скрытых пространств [?, ?, ?]. Предложенные методы позволяют учесть регрессионную компоненту между исходным и целевым сигналами, а также авторегрессионную компоненту целевого сигнала.

Методы снижения размерности пространства понижают размерность исходного пространства объектов, и, как следствие, сложность модели существенно снижается [?, ?, ?]. Алгоритмы снижения размерности находят оптимальные комбинации исходных признаков. Если число таких комбинаций существенно меньше, чем число исходных признаков, то полученное представление снижает размерность. Цель снижения размерности — получение наиболее репрезентативных и информативных комбинаций признаков для решения задачи.

Выбор признаков является частным случаем снижения размерности пространства [?, ?]. Найденные комбинации признаков являются подмножеством исходных признаков. Таким образом отсеиваются шумовые неинформативные признаки. Рассматриваются два типа методов выбора признаков [?, ?, ?]. Первый тип методов не зависит от последующей прогностической модели. Признаки отбираются на основе свойств исходных пространств, а не на основе свойств модели. Второй тип методов отбирает признаки с учётом знания о прогностической модели.

После нахождения оптимального представления данных с помощью снижения размерности, ставится задача нахождения оптимальной метрики в скрытом пространстве объектов [?, ?, ?, ?, ?]. В случае евклидова пространства естественным выбором метрики оказывается квадратичная норма. Задача метрического

обучения заключается в нахождении оптимальной метрики, связывающей объекты.

В качестве прикладной задачи анализируется задача построения нейрокомпьютерного интерфейса [**?**, **?**]. Цель состоит в извлечении информации из сигналов мозговой активности [**?**, **?**, **?**]. В качестве исходных сигналов выступают сигналы электроэнцефалограммы или электрокортикограммы. Целевым сигналом является траектория движения конечности индивидуума. Задача модели построить адекватную и эффективную модель декодирования исходного сигнала в целевой сигнал. Пространство частотных характеристик мозговых сигналов и авторегрессионное пространство целевых сигналов являются чрезвычайно избыточными [**?**, **?**]. Построение модели без учёта имеющихся зависимостей приводит к неустойчивости модели.

В диссертации решается задача декодирования с векторной целевой переменной. Для построения оптимальной модели декодирования сигналов предлагаются методы выбора согласованных моделей с проекцией в скрытое пространство. Исходные и целевые сигналы проецируются в пространство существенно меньшей размерности. Для связи проекций исходного и целевого сигнала предлагаются методы согласования. Рассматриваются гетерогенные наборы сигналов, природа источников измерений различны. Рассматриваются как линейные методы декодирования, так и их нелинейные обобщения. Доказаны теоремы об оптимальности предложенных методов выбора моделей.

Цели работы.

1. Исследовать свойства решения задачи декодирования сигналов с векторной целевой переменной.
2. Предложить методы снижения размерности пространства, учитывающие зависимости как в пространстве исходных сигналов, так и в целевом пространстве.
3. Предложить процедуру выбора признаков для задачи декодирования сигналов.
4. Исследовать свойства линейных и нелинейных моделей для решения поставленной модели. Получить теоретические оценки оптимальности моделей.
5. Провести вычислительные эксперименты для проверки адекватности предложенных методов.

Основные положения, выносимые на защиту.

1. Исследована проблема снижения размерности сигналов в коррелированных пространствах высокой размерности. Предложены методы декодирования сигналов, учитывающие зависимости как в исходном, так и в целевом пространстве сигналов.
2. Доказаны теоремы об оптимальности предлагаемых методов декодирования сигналов. Предлагаемые методы выбирают согласованные модели в случае избыточной размерности описания данных.

3. Предложены методы выбора признаков, учитывающие зависимости как в исходном, так и в целевом пространстве. Предложенные методы доставляют устойчивые и адекватные решения в пространствах высокой размерности.
4. Предложены нелинейные методы согласования скрытых пространств для данных со сложноорганизованной целевой переменной. Предложен метод выбора наиболее релевантных параметров для оптимизации нелинейной модели. Исследованы свойства предлагаемого метода.
5. Предложен алгоритм метрического обучения для временных рядов с процедурой их выравнивания.
6. Предложен ряд моделей для прогнозирования гетерогенных наборов сигналов для задачи построения нейрокомпьютерных интерфейсов. Проведены вычислительные эксперименты, подтверждающие адекватность моделей.

Методы исследования. Для достижения поставленных целей используются линейные и нелинейные методы регрессионного анализа. Для анализа временных рядов используются классические авторегрессионные методы. Для извлечения признаков используются частотные характеристики временного ряда. Для построения скрытого пространства используются линейные методы снижения размерности пространства, их нелинейные модификации, а также нейросетевые методы. Для выбора признаков наряду с классическими методами, используются методы, основанные на решении задачи квадратичного программирования. Для построения метрического пространства используются методы условной выпуклой оптимизации.

Научная новизна. Предложены методы построения моделей декодирования сигналов, учитывающие структуры пространств исходных и целевых переменных. Предложены методы проекции сигналов в скрытое пространство, а также процедуры согласования образов. Предложены методы выбора признаков с помощью квадратичного программирования. Предложен метод выбора параметров нелинейной модели для оптимизации с помощью выбора признаков. Предложены методы построения оптимального метрического пространства для задачи анализа временных рядов.

Теоретическая значимость. Доказаны теоремы об оптимальности предлагаемых моделей декодирования сигналов. Доказаны теоремы о корректности рассматриваемых согласованных моделей проекций в скрытое пространство. Доказаны теоремы о достижении точки равновесия для предлагаемых методов выбора признаков.

Практическая значимость. Предложенные в работе методы предназначены для декодирования множества временных рядов сигналов электрокортикограмм, а также нестационарных временных рядов; выбора оптимальных

частотных характеристик сигналов; выбора наиболее информативных параметров модели; классификации и кластеризации временных рядов физической активности.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой результатов предлагаемых методов на реальных данных, публикациями результатов в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. Р. В. Исаченко. Метрическое обучение в задачах мультиклассовой классификации временных рядов. *Международная научная конференция «Ломоносов»*, 2016.
2. R. G. Neychev, A. P. Motrenko, R. V. Isachenko, A. S. Inyakin, and V. V. Strijov. Multimodel forecasting multiscale time series in internet of things. *Международная научная конференция «11th International Conference on Intelligent Data Processing: Theory and Applications»*, 2016.
3. Р. В. Исаченко, И. Н. Жариков, и А. М. Бочкарёв. Локальные модели для классификации объектов сложной структуры. *Всероссийская научная конференция «Математические методы распознавания образов»*, 2017.
4. R. V. Isachenko and V. V. Strijov. Dimensionality reduction for multicorrelated signal decoding with projections to latent space. *Международная научная конференция «12th International Conference on Intelligent Data Processing: Theory and Applications»*, 2018.
5. Р. В. Исаченко, В. В. Стрижов. Снижение размерности в задаче декодирования временных рядов. *Международная научная конференция «13th International Conference on Intelligent Data Processing: Theory and Applications»*, 2020.

Работа поддержана грантами Российского фонда фундаментальных исследований.

1. 19-07-00885, Российский фонд фундаментальных исследований в рамках гранта «Выбор моделей в задачах декодирования временных рядов высокой размерности».
2. 16-37-00485, Российский фонд фундаментальных исследований в рамках гранта «Развитие методов выбора признаков в условиях мультиколлинеарности».
3. 16-07-01160, Российский фонд фундаментальных исследований в рамках гранта «Развитие теории обучения по предпочтениям с использованием частично упорядоченных множеств экспертных оценок».
4. 16-07-01154, Российский фонд фундаментальных исследований в рамках гранта «Новые методы прогнозирования на базе субквадратичного анализа метрических конфигураций».

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 6 печатных изданиях, 5 из которых изданы в журналах, рекомендованных ВАК.

Структура и объем работы. Диссертация состоит из оглавления, введения, 6 глав, заключения, списка иллюстраций, списка таблиц, списка основных обозначений и списка литературы из ВСТАВИТЬ ЧИСЛО наименований. Основной текст занимает ВСТАВИТЬ ЧИСЛО страниц.

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

Основное содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулированы цели и методы исследования, обоснована научная новизна, теоретическая и практическая значимости полученных результатов.

Глава 1. Постановка задачи декодирования сигналов.

В данной главе ставится общая задача декодирования временных рядов. Приводится обзор стандартных методов анализа временных рядов. Ставится задача построения оптимальной линейной регрессионной модели декодирования. Приведен обзор широко используемых методов снижения размерности пространства, их обобщений и модификаций.

Регрессионная модель в пространстве высокой размерности.

Пусть $\mathbb{X} \subset \mathbb{R}^n$ — пространство исходной переменной, $\mathbb{Y} \subset \mathbb{R}^r$ — пространство целевой переменной. Пусть задано множество объектов $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$, где $\mathbf{x}_i \in \mathbb{X}$ — исходный объект, $\mathbf{y} \in \mathbb{Y}$ — целевой объект.

Обозначим за $\mathbf{X} \in \mathbb{R}^{m \times n}$ матрицу исходной переменной, за $\mathbf{Y} \in \mathbb{R}^{n \times k}$ матрицу целевой переменной:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\intercal = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n]; \quad \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]^\intercal = [\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_r].$$

Столбцы $\boldsymbol{\chi}_j, j = 1, \ldots, n$ матрицы $\mathbf{X}$ являются признаками исходного объекта, столбцы $\boldsymbol{\nu}_j, j = 1, \ldots, r$ матрицы $\mathbf{Y}$ являются целевыми векторами.

Предполагается, что между исходным объектом $\mathbf{x}$ и целевым объектом $\mathbf{y}$ существует зависимость. Требуется построить прогностическую модель $\mathbf{f} : \mathbb{X} \to \mathbb{Y}$ из пространства исходной переменной в пространство целевой переменной.

Задача восстановления регрессионной зависимости состоит в нахождении оптимальной модели $\mathbf{f}^*$ по заданным матрицам $\mathbf{X}$ и $\mathbf{Y}$. Под оптимальностью понимается нахождение такой модели, которая бы доставляла минимум некоторой функции ошибки $\mathcal{L}$:

$$\mathbf{f}^* = \arg\min_{\mathbf{f}} \mathcal{L}(\mathbf{f}, \mathbf{X}, \mathbf{Y}). \tag{1}$$

Задача поиска оптимальной модели является задачей функциональной оптимизации. Для сужения пространства поиска моделей будем рассматривать параметрические модели $\mathbf{f}(\mathbf{x}, \boldsymbol{\Theta})$, где $\boldsymbol{\Theta}$ — *параметры модели*. Таким образом между объектами $\mathbf{x}$ и $\mathbf{y}$ существует зависимость вида

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \boldsymbol{\Theta}) + \boldsymbol{\varepsilon}, \tag{2}$$

где $\mathbf{f}$ — параметрическая прогностическая модель, $\boldsymbol{\Theta}$ — параметры модели, $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ — вектор регрессионных остатков.

Задача (1) сводится к задаче поиска набора оптимальных параметров

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y}). \tag{3}$$

В диссертации рассматривается случай избыточной размерности пространств $\mathbb{X}$, $\mathbb{Y}$. В таком случае решение задачи (3) оказывается неустойчивым. Рассмотрим в качестве примера задачу восстановления линейной регрессии.

Предположим, что зависимость $\mathbf{f}(\mathbf{x}, \boldsymbol{\Theta})$ линейная:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \boldsymbol{\Theta}) + \boldsymbol{\varepsilon} = \boldsymbol{\Theta}^{\intercal}\mathbf{x} + \boldsymbol{\varepsilon}, \tag{4}$$

где $\boldsymbol{\Theta} \in \mathbb{R}^{n \times r}$ — матрица параметров модели.

Оптимальные параметры $\boldsymbol{\Theta}$ определяются минимизацией функции ошибки $\mathcal{L}(\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y})$. При решении задачи линейной регрессии в качестве такой функции ошибки рассматривается квадратичная функция потерь:

$$\mathcal{L}(\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y}) = \left\| \underset{m \times r}{\mathbf{Y}} - \underset{m \times n}{\mathbf{X}} \cdot \underset{r \times n}{\boldsymbol{\Theta}} \right\|_2^2 \to \min_{\boldsymbol{\Theta}}. \tag{5}$$

Решением (5) является следующая матрица:

$$\boldsymbol{\Theta} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{Y}.$$

Наличие линейной зависимости между столбцами матрицы $\mathbf{X}$ приводит к неустойчивому решению задачи оптимизации (5). Если существует вектор $\boldsymbol{\alpha} \neq \mathbf{0}_n$ такой, что $\mathbf{X}\boldsymbol{\alpha} = \mathbf{0}_m$, то добавление $\boldsymbol{\alpha}$ к любому столбцу матрицы $\boldsymbol{\Theta}$ не меняет значение функции потерь $\mathcal{L}(\boldsymbol{\Theta}, \mathbf{X}, \mathbf{Y})$. В этом случае матрица $\mathbf{X}^{\intercal}\mathbf{X}$ близка к сингулярной и не обратима. Чтобы избежать сильной линейной зависимости между признаками, в данной работе исследуются методы снижения размерности и выбора признаков.

Задача декодирования сигналов.

Задача декодирования сигналов состоит в восстановлении регрессионной зависимости (1) между наборами гетерогенных сигналов.

Пусть имеется два множества временных рядов $\mathcal{S}_{\mathbf{x}} = \{\mathbf{s}_{\mathbf{x}}^i\}_{i=1}^m$ и $\mathcal{S}_{\mathbf{y}} = \{\mathbf{s}_{\mathbf{y}}^i\}_{i=1}^r$, состоящие из m и r временных рядов соответственно. Первое множество $\mathcal{S}_{\mathbf{x}}$

является множеством временных рядов m исходных сигналов. Второе множество $\mathcal{S}_{\mathbf{y}}$ является множеством временных рядов r целевых сигналов. Каждый временной ряд $\mathbf{s} = (s_1, s_2, \ldots, s_T)$ является последовательностью измерений некоторый величины в течение времени.

Определение 1. *Временное представление* $\mathbf{x}_t = ([\mathbf{s}_{\mathbf{x}}^1]_t, \ldots, [\mathbf{s}_{\mathbf{x}}^m]_t) \in \mathbb{R}^m$ состоит из измерений временных рядов исходных сигналов в момент времени t. Аналогично временное представление $\mathbf{y}_t = ([\mathbf{s}_{\mathbf{y}}^1]_t, \ldots, [\mathbf{s}_{\mathbf{y}}^r]_t) \in \mathbb{R}^r$ состоит из измерений временных рядов целевых сигналов в момент времени t.

Определение 2. Определим *представление предыстории* длины h для момента времени t множества временных рядов исходных сигналов $\mathcal{S}_{\mathbf{x}}$ как совокупность представлений $\mathbf{X}_{t,h} = [\mathbf{x}_{t-h+1}, \ldots, \mathbf{x}_t]^{\mathsf{T}} \in \mathbb{R}^{h \times m}$. Аналогично определим представление предыстории длины h для момента времени t множества временных рядов целевых сигналов $\mathcal{S}_{\mathbf{y}}$ как совокупность представлений $\mathbf{Y}_{t,h} = [\mathbf{y}_{t-h+1}, \ldots, \mathbf{y}_t]^{\mathsf{T}} \in \mathbb{R}^{h \times r}$.

Определение 3. Определим *представление горизонта прогнозирования* длины p для момента времени t множества временных рядов исходных сигналов $\mathcal{S}_{\mathbf{x}}$ как совокупность представлений $\mathbf{X}_{t,p} = [\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+p}]^{\mathsf{T}} \in \mathbb{R}^{p \times m}$. Аналогично определим представление горизонта прогнозирования длины p для момента времени t множества временных рядов целевых сигналов $\mathcal{S}_{\mathbf{y}}$ как совокупность представлений $\mathbf{Y}_{t,r} = [\mathbf{y}_{t+1}, \ldots, \mathbf{y}_{t+p}]^{\mathsf{T}} \in \mathbb{R}^{p \times r}$.

Задача авторегрессионного декодирования состоит в построении прогностической модели $\mathbf{f}^{\mathrm{AR}}$, дающий прогноз представления горизонта прогнозирования множества временных рядов по представлению предыстории прогнозирования того же множества временных рядов.

Определение 4. Прогностическая модель $\mathbf{f}_{\mathbf{x}}^{\mathrm{AR}} : \mathbb{R}^{h \times m} \to \mathbb{R}^{p \times m}$ является *авторегрессионной моделью*, которая по представлению предыстории $\mathbf{X}_{t,h}$ множества временных рядов исходных сигналов $\mathcal{S}_{\mathbf{x}}$ предсказывает представление горизонта прогнозирования $\mathbf{X}_{t,p}$ множества временных рядов исходных сигналов $\mathcal{S}_{\mathbf{x}}$. Аналогично вводится прогностическая модель $\mathbf{f}_{\mathbf{y}}^{\mathrm{AR}} : \mathbb{R}^{h \times r} \to \mathbb{R}^{p \times r}$ для множества целевых сигналов $\mathcal{S}_{\mathbf{y}}$.

Суть авторегрессионного декодирования заключается в предсказании будущего прогноза сигнала по его же предыстории.

Определение 5. Определим задачу *регрессионного декодирования* как задачу построения прогностической модели $\mathbf{f}_{\mathbf{xy}}^{\mathrm{R}} : \mathbb{R}^{h \times m} \to \mathbb{R}^{p \times r}$, которая по представлению предыстории $\mathbf{X}_{t,h}$ множества временных рядов исходных сигналов $\mathcal{S}_{\mathbf{x}}$ предсказывает представление горизонта прогнозирования $\mathbf{Y}_{t,p}$ множества временных рядов целевых сигналов $\mathcal{S}_{\mathbf{y}}$.

Отличие регрессионного декодирования от авторегрессионного декодирования состоит в том, что в случае регрессионного декодирования представление предыстории и представление горизонта прогнозирования получены из временных рядов разных пространств. Таким образом предыстория получена из множества исходных сигналов, в то время как горизонт прогнозирования получен из множества целевых сигналов. Пространства исходных и целевых сигналов могут являться существенно гетерогенными и обладать разными свойствами.

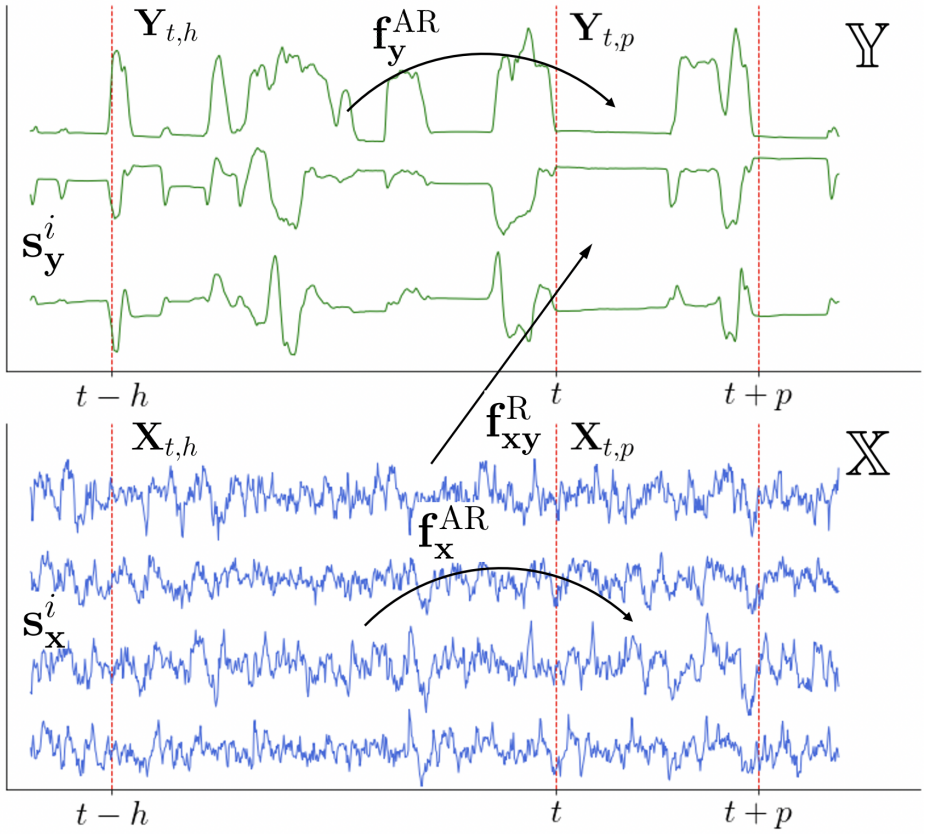

Рис. 1: Схема построения моделей декодирования

Определение 6. Общая *задача декодирования* состоит в построении прогностической модели $\mathbf{f}_{\mathbf{xy}} : \mathbb{R}^{h_x \times m} \times \mathbb{R}^{h_y \times r} \to \mathbb{R}^{p \times r}$, которая по представлениям предыстории $\mathbf{X}_{t,h_x}$ и $\mathbf{Y}_{t,h_y}$ временных рядов исходных и целевых сигналов предсказывает представление горизонта прогнозирования $\mathbf{Y}_{t,r}$ временных рядов целевых сигналов.

Отметим, что авторегрессионная модель $\mathbf{f}^{\mathrm{AR}}_{\mathbf{y}}$ и регрессионная модель $\mathbf{f}^{\mathrm{R}}_{\mathbf{xy}}$ являются частными случаями общей задачи декодирования. А именно, авторегрессионная модель $\mathbf{f}^{\mathrm{AR}}_{\mathbf{y}}$ соответствует случаю пустой предыстории временных рядов исходных сигналов (случаю $h_x = 0$), а регрессионная модель $\mathbf{f}^{\mathrm{R}}_{\mathbf{xy}}$ соответствует случаю пустой предыстории временных рядов целевых сигналов (случаю $h_y = 0$).

На Рис. 1 схематично продемонстрированы принципы построения введенных моделей декодирования временных рядов.

Для построения авторегрессионной модели декодирования временных рядов широко используются два класса линейных методов: авторегрессионные модели и модели скользящего среднего [?, ?]. Авторегрессионные модели AR(p) строят прогноз в виде линейной комбинации p предыдущих значений временного ряда. Модели скользящего среднего MA(q) вместо предыдущих значений временного ряда используют комбинацию ошибок. Модель ARMA(p, q) [?] является комбинацией двух описанных подходов. ARMA(p, q) задает модель как линейную комбинацию p предыдущих значений временного ряда и q предыдущих значе-

ний ошибок. Для нахождения оптимальных параметров p и q модели ARMA используются автокорреляционная и частная автокорреляционная функции.

Модель ARMA используется для стационарных временных рядов, отвечающим строгим статистическим предположениям. На практике встречается огромное количество нестационарных временных рядов подверженных тренду, сезонности или цикличности. Модель ARIMA(p, d, q) [**?**] обобщает модель ARMA для случая нестационарных временных рядов. ARIMA берёт разности порядка d от исходного временного ряда для достижения стационарности данных. При этом на практике оказывается достаточным положить $d = 1$. Заметим, что при $d = 0$ модель ARIMA эквивалентна модели ARMA. Полезным обобщением модели ARIMA является модель AFRIMA [**?**]. Модель позволяет задать параметр d в виде вещественного числа.

Модель ARIMA плохо справляется с сезонными временными рядами. В работе [**?**] была предложена модель SARIMA, которая вводит в модель учет сезонной компоненты.

Задача декодирования временных рядов декомпозируется на следующие подзадачи.

- Порождение признакового пространства. Данный этап включает в себя процедуру извлечения признаков из исходных значений сигналов. Процедура порождения признакового пространства может быть основана на экспертных знаниях или же являться моделью машинного обучения. Данная подзадача подробно рассмотрена в главе 0.3.
- Снижение размерности пространства или выбор признаков. Исходные временные ряды, а также порожденное признаковое пространство оказывается избыточным, что приводит к избыточности и неустойчивости модели. Методы снижения размерности и выбора признаков подробно изложены в главах **??** и 0.1.
- Построение модели. После нахождения оптимального низкоразмерного представления исходных данных ставится задача выбора оптимальной модели декодирования.

Обзор методов снижения размерности для задачи декодирования.

Методы снижения размерности позволяют найти низкоразмерное представление исходных данных. Найденное представление используется для построения прогностической модели. При этом метод снижения размерности может учитывать как зависимости в исходном объекте $\mathbf{x}$, так и в целевом объекте $\mathbf{y}$.

Метод главных компонент для задачи декодирования. Для устранения линейной зависимости и снижения размерности исходного пространства объектов широко используется метод главных компонент (principal component analysis, PCA). Метод PCA находит низкоразмерное представление матрицы $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathsf{T}}$, такое что новое представление $\mathbf{T} \in \mathbb{R}^{m \times l}$ содержит максимальную долю дисперсии исходной матрицы. При этом матрица отображения $\mathbf{P} \in \mathbb{R}^{n \times l}$

является ортогональной ($\mathbf{P}^\intercal\mathbf{P} = \mathbf{I}$) и содержит правые собственные вектора матрицы ковариаций $\mathbf{X}^\intercal\mathbf{X}$.

Метод PCA является базовым методом снижения размерности пространства. Существует множество модификаций базового метода. Вероятностный PCA [**?**] рассматривает задачу снижения размерности в терминах вероятностной модели, решая задачу с помощью вариационного EM алгоритма. Разреженный PCA [**?**] вводит в постановку задачи lasso регуляризацию для того, чтобы сделать матрицу отображения $\mathbf{P}$ разреженной и более интерпретируемой. Нелинейный ядерный PCA [**?**] отображает исходные данные с помощью нелинейного отображения и использует RKHS для решения исходной задачи.

После нахождения матрицы отображения $\mathbf{P}$ задача (5) принимает вид

$$\mathcal{L}(\mathbf{B}, \mathbf{T}, \mathbf{Y}) = \left\| \underset{m\times r}{\mathbf{Y}} - \underset{m\times l}{\mathbf{T}} \cdot \underset{l\times r}{\mathbf{B}} \right\|_2^2 \to \min_{\mathbf{B}}. \tag{6}$$

Модель прогнозирования (4) в случае снижения размерности с помощью PCA принимает вид:

$$\mathbf{y} = \mathbf{B}\mathbf{t} + \boldsymbol{\varepsilon} = \mathbf{B}\mathbf{P}\mathbf{x} + \boldsymbol{\varepsilon} = \boldsymbol{\Theta}\mathbf{x} + \boldsymbol{\varepsilon}, \text{ где } \boldsymbol{\Theta} = \mathbf{B}\mathbf{P}. \tag{7}$$

Метод частичных наименьших квадратов для задачи декодирования. Основным недостатком метода PCA является отсутствие учёта взаимосвязи между исходными признаками $\boldsymbol{\chi}_j$ и целевыми векторами $\boldsymbol{\nu}_j$. Метод частичных наименьших квадратов (partial least squares, PLS) проецирует матрицу объектов $\mathbf{X}$ и матрицу ответов $\mathbf{Y}$ в скрытое пространство малой размерностью l ($l < n$). Метод PLS находит в скрытом пространстве матрицы $\mathbf{T}, \mathbf{U} \in \mathbb{R}^{m\times l}$, которые лучше всего описывают исходные матрицы $\mathbf{X}$ и $\mathbf{Y}$. При этом PLS максимизирует ковариацию между столбцами матриц $\mathbf{T}$ и $\mathbf{U}$ соответственно.

Метод PLS был впервые предложен в работах [**?**, **?**, **?**]. Подробное описание алгоритма приведено в работах [**?**, **?**, **?**, **?**, **?**]. В работах [**?**, **?**] приведен обзор обобщений базовой модели PLS. В работе [**?**] приведена модификация метода PLS для получения разреженного набора признаков.

Матрица исходных объектов $\mathbf{X}$ и матрица целевых объектов $\mathbf{Y}$ проецируются на скрытое пространство следующим образом:

$$\underset{m\times n}{\mathbf{X}} = \underset{m\times l}{\mathbf{T}} \cdot \underset{l\times n}{\mathbf{P}^\intercal} + \underset{m\times n}{\mathbf{E}_\mathbf{x}} = \sum_{k=1}^{l} \underset{m\times 1}{\boldsymbol{\tau}_k} \cdot \underset{1\times n}{\mathbf{p}_k^\intercal} + \underset{m\times n}{\mathbf{E}_\mathbf{x}}, \tag{8}$$

$$\underset{m\times r}{\mathbf{Y}} = \underset{m\times l}{\mathbf{U}} \cdot \underset{l\times r}{\mathbf{Q}^\intercal} + \underset{m\times r}{\mathbf{E}_\mathbf{y}} = \sum_{k=1}^{l} \underset{m\times 1}{\boldsymbol{\nu}_k} \cdot \underset{1\times r}{\mathbf{q}_k^\intercal} + \underset{m\times r}{\mathbf{E}_\mathbf{y}}. \tag{9}$$

Здесь $\mathbf{T}$ и $\mathbf{U}$ — образы исходных матриц в скрытом пространстве, причём столбцы матрицы $\mathbf{T}$ ортогональны; $\mathbf{P}$ и $\mathbf{Q}$ — матрицы перехода; $\mathbf{E}_\mathbf{x}$ и $\mathbf{E}_\mathbf{y}$ —

матрицы остатков. Метод PLS максимизирует линейную зависимость между столбцами матриц $\mathbf{T}$ и $\mathbf{U}$

$$\mathbf{U} \approx \mathbf{TB}, \quad \mathbf{B} = \operatorname{diag}(\beta_k), \quad \beta_k = \boldsymbol{\nu}_k^{\intercal}\boldsymbol{\tau}_k/(\boldsymbol{\tau}_k^{\intercal}\boldsymbol{\tau}_k),$$

где $\{\boldsymbol{\tau}\}_{i=1}^{l}$, $\{\boldsymbol{\nu}\}_{i=1}^{l}$ — столбцы матриц $\mathbf{T}$ и $\mathbf{U}$ соответственно.

Метод решает следующую оптимизационную задачу:

$$\max_{\|\mathbf{p}\|_2=\|\mathbf{q}\|_2=1} [\operatorname{cov}(\mathbf{Xp}, \mathbf{Yq})^2] = \max_{\mathbf{p},\mathbf{q}} \frac{\mathbf{p}^{\intercal}\mathbf{X}^{\intercal}\mathbf{Yq}}{\sqrt{\mathbf{p}^{\intercal}\mathbf{p}}\sqrt{\mathbf{q}^{\intercal}\mathbf{q}}}. \tag{10}$$

Детальное описание метода PLS с доказательством его корректности приведено в разделе .

Для демонстрации разницы между методами PCA, PLS был проведен модельный эксперимент для случая, когда размерности пространств исходных объектов, целевых объектов и скрытого пространства равны 2 ($n = r = l = 2$). На Рис. **??** показаны результаты работы методов. Синими и зелёными точками изображены исходные объекты $\mathbf{x}_i$ и целевые объекты $\mathbf{y}_i$. Точки $\mathbf{X}$ сгенерированы из нормального распределения с нулевым матожиданием. Точки $\mathbf{Y}$ линейным образом зависят от второй главной компоненты pc_2 матрицы $\mathbf{X}$ и не зависят от первой главной компоненты pc_1. Красным контуром показаны линии уровня матриц ковариаций распределений. Черным изображены единичные окружности. Красные стрелки соответствуют главным компонентам матриц $\mathbf{X}$ и $\mathbf{Y}$. Черные стрелки соответствуют векторам матриц $\mathbf{W}$ и $\mathbf{C}$ метода PLS. Данные матрицы содержат вектора, являющиеся аналогами главных компонент метода PCA. Учёт взаимной связи между матрицами $\mathbf{X}$ и $\mathbf{Y}$ отклоняет вектора $\mathbf{w}_k$ и $\mathbf{c}_k$ от направления главных компонент.

При снижении размерности пространств до одного признака метод PCA выберет первую главную компоненту pc_1, отбросив компоненту pc_2, так как первая компонента объясняет большую часть дисперсии исходной матрицы $\mathbf{X}$. При этом матрица $\mathbf{Y}$ не зависит от pc_1. Тем самым финальная модель окажется не адекватной. Метод PLS позволяет побороться с данной проблемой.

Канонический анализ корреляций для задачи декодирования. Канонический корреляционный анализ (canonical correlation analysis, CCA) широко применяется для поиска взаимосвязи между двумя наборами переменных [**?**, **?**]. Оптимизационная задача CCA похожа на оптимизационную задачу PLS (10) с той лишь разницей, что вместо максимизации ковариации CCA максимизирует корреляцию:

$$\max_{\|\mathbf{p}\|_2=\|\mathbf{q}\|_2=1} [\operatorname{corr}(\mathbf{Xp}, \mathbf{Yq})^2] = \max_{\mathbf{p},\mathbf{q}} \frac{\mathbf{p}^{\intercal}\mathbf{X}^{\intercal}\mathbf{Yq}}{\sqrt{\mathbf{p}^{\intercal}\mathbf{X}^{\intercal}\mathbf{Xp}}\sqrt{\mathbf{q}^{\intercal}\mathbf{Y}^{\intercal}\mathbf{Yq}}}. \tag{11}$$

На Рис. **??** показан результат работы метода. Основное различие состоит в том, что вектора $\mathbf{c}_1$ и $\mathbf{c}_2$ в данном случае становятся ортогональными.

Линейная регрессия	PCA	PLS	CCA
0.01	0.24	0.13	0.13

Таблица 1: Средняя квадратичная ошибка на модельном примере для методов линейной регрессии, PCA, PLS, CCA

В таблице 1 приведены значения квадратичной ошибки $\mathcal{L}(\mathbf{\Theta}, \mathbf{X}, \mathbf{Y})$ для методов линейной регрессии, PCA, PLS и CCA. Линейная регрессия отлично справляется с данной задачей. Ошибка метода PCA наибольшая, что подтверждает факт, что для данной ситуации метод не находит нужных зависимостей в пространстве целевой переменной. Методы PLS и CCA показывают схожие результаты.

Нелинейный ядерный CCA [?, ?, ?, ?] является обобщением базового метода. CCA и ядерный CCA широко используются для задач обучения без учителя [?, ?]. Метод имеет область применения от анализа хемометрических [?] и биологических [?] данных до обработки естественного языка [?, ?], аудиосигналов [?, ?] и компьютерного зрения [?].

В работе [?] впервые было предложено обобщение метода CCA, работающего с нейросетями. Предложенный метод DeepCCA максимизирует корреляцию между представлениями, полученными на выходе нейросети:

$$\max_{\|\mathbf{p}\|_2=\|\mathbf{q}\|_2=1} \left[\operatorname{corr}(\mathbf{g_x}(\mathbf{X}, \mathbf{W_x}) \cdot \mathbf{p}, \mathbf{g_y}(\mathbf{Y}, \mathbf{W_y}) \cdot \mathbf{q})^2\right] = \\ = \max_{\mathbf{p},\mathbf{q}} \frac{\mathbf{p}^\intercal \mathbf{g_x}(\mathbf{X}, \mathbf{W_x})^\intercal \mathbf{g_y}(\mathbf{Y}, \mathbf{W_y})\mathbf{q}}{\sqrt{\mathbf{p}^\intercal \mathbf{g_x}(\mathbf{X}, \mathbf{W_x})^\intercal \mathbf{g_x}(\mathbf{X}, \mathbf{W_x})\mathbf{p}}\sqrt{\mathbf{q}^\intercal \mathbf{g_y}(\mathbf{Y}, \mathbf{W_y})^\intercal \mathbf{g_y}(\mathbf{Y}, \mathbf{W_y})\mathbf{q}}}. \quad (12)$$

Здесь $\mathbf{g_x}(\mathbf{X}, \mathbf{W_x})$ и $\mathbf{g_y}(\mathbf{Y}, \mathbf{W_y})$ — нелинейные проекции исходных и целевых объектов. В статье [?] приведен обширный обзор модификаций нейросетевого CCA для работы с многовидовыми данными. С использованием нейросетевых функций модель декодирования способна учитывать существенно нелинейные зависимости как в исходном пространстве, так и в целевом пространстве. Главным недостатком нейросетевого CCA является вычислительная сложность. В работе [?] предложена релаксация исходной функции потерь, которая способна масштабироваться под работу с большими глубокими моделями нейросетей.

Тензорные линейные методы для задачи декодирования. Если исходный объект $\mathbf{x}$ является не вектором, а тензором более высокого порядка, то для построения модели тензор может быть вытянут в вектор [?]. В таком случае модель не учитывает имеющиеся зависимости между различными направлениями исходного тензора. Для учета таких зависимостей используются тензорные версии метода PLS [?, ?, ?].

Многомодальные данные в задаче декодирования. Исходный объект может иметь несколько модальностей. Примерами таких модальностей могут

быть выровненные аудио и видео [?, ?], аудио и артикуляция [?], изображение и текстовая аннотация [?, ?, ?], параллельный корпус текстов [?, ?, ?, ?].

В случае если для каждого объекта имеется более двух модальностей, то для построения скрытого пространства применяются два класса подходов. Первый подход состоит в построении скрытого пространства для каждой пары модальностей объекта [?, ?]. Второй же подход состоит в построении общего единого скрытого пространства для всех модальностей [?, ?].

Глава 2. Задача построения согласованных моделей декодирования

В данной главе приводится формальная постановка задачи декодирования в терминах проекций в скрытое пространство. Вводятся понятия скрытого пространства и процедуры согласования образов. Доказываются теоремы о выборе оптимальной модели декодирования.

Процесс согласования моделей в пространстве высокой размерности.

Для постановки задачи декодирования введём предположения о структурах пространств $\mathbb{X}$ и $\mathbb{Y}$.

Предположение 1. Рассмотрим случай, когда пространства $\mathbb{X}$ и $\mathbb{Y}$ имеют избыточную размерность. Это означает, что объекты $\mathbf{x}$ и $\mathbf{y}$ принадлежат некоторым многообразиям низкой размерности. В простейшем случае такие многообразия могут являться вложениями или линейными подпространствами.

Определение 7. Назовём пространство $\mathbb{T} \subset \mathbb{R}^l$ *скрытым пространством* для пространства $\mathbb{X} \subset \mathbb{R}^n$ ($l \leqslant n$), если существуют функция $\boldsymbol{\varphi}_{\mathbf{x}} : \mathbb{X} \to \mathbb{T}$ и функция $\boldsymbol{\psi}_{\mathbf{x}} : \mathbb{T} \to \mathbb{X}$ такие что

$$\text{для любого } \mathbf{x} \in \mathbb{X} \quad \text{существует } \mathbf{t} \in \mathbb{T} : \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x})) = \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{t}) = \mathbf{x}.$$

Функцию $\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x})$ назовём *функцией кодирования* объекта $\mathbf{x}$, функцию $\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{t})$ назовём *функцией декодирования.*

Аналогично введём определение *скрытого пространства* $\mathbb{U} \subset \mathbb{R}^s$ для целевого пространства $\mathbb{Y}$, *функции кодирования* $\boldsymbol{\varphi}_{\mathbf{y}} : \mathbb{Y} \to \mathbb{U}$ и *декодирования* $\boldsymbol{\psi}_{\mathbf{y}} : \mathbb{U} \to \mathbb{Y}$ такие что

$$\text{для любого } \mathbf{y} \in \mathbb{Y} \quad \text{существует } \mathbf{u} \in \mathbb{U} : \boldsymbol{\psi}_{\mathbf{y}}(\boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{y})) = \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{u}) = \mathbf{y}.$$

Образы матрицы исходных объектов $\mathbf{X}$ и матрицы целевых объектов $\mathbf{Y}$ в скрытых пространствах $\mathbb{T}$ и $\mathbb{U}$ имеют вид

$$\mathbf{T} = \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{X}) = [\mathbf{t}_1, \ldots, \mathbf{t}_m]^{\mathsf{T}} = [\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_l],$$
$$\mathbf{U} = \boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{U}) = [\mathbf{u}_1, \ldots, \mathbf{u}_m]^{\mathsf{T}} = [\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_s].$$

Здесь строки $\{\mathbf{t}_i\}_{i=1}^m$ матрицы $\mathbf{T}$ и строки $\{\mathbf{u}_i\}_{i=1}^m$ матрицы $\mathbf{U}$ являются образами исходных объектов $\{\mathbf{x}_i\}_{i=1}^m$ и целевых объектов $\{\mathbf{y}_i\}_{i=1}^m$. Столбцы $\{\boldsymbol{\tau}_j\}_{j=1}^l$ матрицы $\mathbf{T}$ и столбцы $\{\boldsymbol{\nu}_j\}_{j=1}^s$ матрицы $\mathbf{U}$ являются скрытыми векторами.

Определение 8. Будем говорить, что скрытые пространства $\mathbb{T}$ и $\mathbb{U}$ являются *согласованными*, если существует *функция связи* $\mathbf{h} : \mathbb{T} \to \mathbb{U}$, такая что

$$\text{для любого } \mathbf{u} \in \mathbb{U} \quad \text{существует } \mathbf{t} \in \mathbb{T} : \mathbf{u} = \mathbf{h}(\mathbf{t}).$$

Предположение 2. Предположим, что в задаче прогнозирования (1) пространства $\mathbb{T}$ и $\mathbb{U}$ являются скрытыми для пространств $\mathbb{X}$ и $\mathbb{Y}$ соответственно. Предположим также, что для данных скрытых пространств $\mathbb{T}$ и $\mathbb{U}$ существует функция связи $\mathbf{h} : \mathbb{T} \to \mathbb{U}$. Тогда выполнено

$$\text{для любого } \mathbf{y} \in \mathbb{Y} \quad \text{существует } \mathbf{x} \in \mathbb{X} : \mathbf{y} = \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{u}) = \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{h}(\mathbf{t})) = \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{h}(\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{x}))),$$

и общая схема задачи поиска согласованной модели декодирования принимает вид следующей коммутативной диаграммы:

$$\begin{array}{ccc} \mathbb{X} \subset \mathbb{R}^n & \xrightarrow{\quad\mathbf{f}\quad} & \mathbb{Y} \subset \mathbb{R}^r \\ \boldsymbol{\varphi}_{\mathbf{x}} \big\uparrow\big\downarrow \boldsymbol{\psi}_{\mathbf{x}} & & \boldsymbol{\psi}_{\mathbf{y}} \big\uparrow\big\downarrow \boldsymbol{\varphi}_{\mathbf{y}} \\ \mathbb{T} \subset \mathbb{R}^{\ell} & \xrightarrow{\quad\mathbf{h}\quad} & \mathbb{U} \subset \mathbb{R}^s \end{array} \tag{13}$$

Определение 9. Согласно диаграмме (13), определим *согласованную* модель декодирования $\mathbf{f} : \mathbb{X} \to \mathbb{Y}$ как суперпозицию

$$\mathbf{f} = \boldsymbol{\psi}_{\mathbf{y}} \circ \mathbf{h} \circ \boldsymbol{\varphi}_{\mathbf{x}}. \tag{14}$$

Таким образом задача прогнозирования (1) сводится к поиску согласованной модели декодирования (14). Для поиска оптимальных параметров функций кодирования $\boldsymbol{\varphi}_{\mathbf{x}}$ и $\boldsymbol{\varphi}_{\mathbf{y}}$, декодирования $\boldsymbol{\psi}_{\mathbf{x}}$ и $\boldsymbol{\psi}_{\mathbf{y}}$, а также функции связи $\mathbf{h}$ ставится задача максимизации *функции согласования скрытых векторов*

$$g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \quad g(\boldsymbol{\tau}, \boldsymbol{\nu}) \to \max_{\boldsymbol{\varphi}_{\mathbf{x}}, \boldsymbol{\varphi}_{\mathbf{y}}, \mathbf{h}}.$$

Каждая пара скрытых векторов $\boldsymbol{\tau}, \boldsymbol{\nu}$ ищется последовательно.

Сформулируем примеры методов снижения размерности пространства, описанные в разделе , в терминах задачи построения согласованной модели декодирования.

Метод главных компонент. Метод главных компонент снижает размерность исходных данных и сохраняет максимальную дисперсию между полученными проекциями. Линейная модель PCA представляет собой ортогональное линейное преобразование исходного признакового пространства в скрытое пространство меньшей размерности.

Функции кодирования $\boldsymbol{\varphi}_{\mathbf{x}} : \mathbb{R}^m \to \mathbb{R}^l$ и декодирования $\boldsymbol{\psi}_{\mathbf{x}} : \mathbb{R}^l \to \mathbb{R}^m$ имеют вид

$$\boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{X}) = \underset{n\times m}{\mathbf{X}} \cdot \underset{m\times l}{\mathbf{P}}^{\intercal}, \quad \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{T}) = \underset{n\times l}{\mathbf{T}} \cdot \underset{l\times m}{\mathbf{P}},$$

где $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_l]$. Здесь матрица $\mathbf{P}$ является ортогональной матрицей, то есть $\mathbf{P}^{-1} = \mathbf{P}^\intercal$.

Скрытые вектора $\boldsymbol{\tau}$ строятся так, чтобы выборочная дисперсия столбцов проекций матрицы $\mathbf{X}$ была максимальной:

$$\mathbf{p} = \arg\max_{\|\mathbf{p}\|_2=1} g(\boldsymbol{\tau}) = \arg\max_{\|\mathbf{p}\|_2=1} [\mathrm{var}(\boldsymbol{\tau})] = \arg\max_{\|\mathbf{p}\|_2=1} [\mathrm{var}(\mathbf{Xp})],$$

где $\mathrm{var}(\boldsymbol{\tau})$ — выборочная дисперсия.

Метод PCA не согласует исходные переменные и целевые переменные. А именно метод PCA не находит функции кодирования $\boldsymbol{\varphi}_\mathbf{y}$ и декодирования $\boldsymbol{\psi}_\mathbf{y}$, а также функцию связи $\mathbf{h}$. При этом функция согласования скрытых векторов $g(\boldsymbol{\tau})$ зависит только от одного аргумента. Из-за этого зависимости в обоих пространствах не учитываются. Пример некорректной работы метода в случае наличия зависимостей как в исходном, так и в целевом пространстве, показан в разделе .

Метод наименьших частичных квадратов и канонический анализ корреляций. В методах PLS и CCA функции кодирования и декодирования имеют вид

$$\boldsymbol{\varphi}_\mathbf{x}(\mathbf{X}) = \mathbf{XW}, \quad \boldsymbol{\varphi}_\mathbf{y}(\mathbf{Y}) = \mathbf{YC},$$
$$\boldsymbol{\psi}_\mathbf{x}(\mathbf{T}) = \mathbf{TP}^\intercal, \quad \boldsymbol{\psi}_\mathbf{y}(\mathbf{U}) = \mathbf{UQ}^\intercal.$$

Функция связи $\mathbf{h}$ имеет вид линейной модели, связывающей образы проекций в скрытом пространстве $\mathbf{u} = \mathbf{h}(\mathbf{t}) = \mathbf{B}^\intercal\mathbf{t}$. В данном случае схема декодирования (13) принимает вид следующей коммутативной диаграммы.

$$\begin{array}{ccc} \mathbb{X} \subset \mathbb{R}^n & \xrightarrow{\quad\mathbf{f}\quad} & \mathbb{Y} \subset \mathbb{R}^r \\ \mathbf{W} \uparrow\downarrow \mathbf{P} & & \mathbf{Q} \uparrow\downarrow \mathbf{C} \\ \mathbb{T} \subset \mathbb{R}^\ell & \xrightarrow{\quad\mathbf{B}\quad} & \mathbb{U} \subset \mathbb{R}^s \end{array}$$

В разделе приводится подробная процедура нахождения оптимальных матриц $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{C}$, $\mathbf{B}$ с доказательством корректности.

Различие между методами PLS и CCA заключается в виде функции согласования g. Для метода PLS функция согласования скрытых векторов имеет вид $g(\boldsymbol{\tau}, \boldsymbol{\nu}) = \mathrm{cov}(\boldsymbol{\tau}, \boldsymbol{\nu})$, а для метода CCA: $g(\boldsymbol{\tau}, \boldsymbol{\nu}) = \mathrm{corr}(\boldsymbol{\tau}, \boldsymbol{\nu})$.

Нелинейный канонический анализ корреляций. Помимо линейных моделей декодирования рассматриваются нелинейные методы. В данном случае функции кодирования и декодирования являются нелинейными нейросетями

вида

$$\begin{aligned}
\mathbf{T} &= \boldsymbol{\varphi}_{\mathbf{x}}(\mathbf{X}) = \mathbf{W}_{\mathbf{x}}^{L}\sigma(\dots\sigma(\mathbf{W}_{\mathbf{x}}^{2}\sigma(\mathbf{X}\mathbf{W}_{\mathbf{x}}^{1}))\dots),\\
\mathbf{U} &= \boldsymbol{\varphi}_{\mathbf{y}}(\mathbf{Y}) = \mathbf{W}_{\mathbf{y}}^{L}\sigma(\dots\sigma(\mathbf{W}_{\mathbf{y}}^{2}\sigma(\mathbf{Y}\mathbf{W}_{\mathbf{y}}^{1}))\dots),\\
\mathbf{X} &= \boldsymbol{\psi}_{\mathbf{x}}(\mathbf{T}) = \mathbf{W}_{\mathbf{t}}^{L}\sigma(\dots\sigma(\mathbf{W}_{\mathbf{t}}^{2}\sigma(\mathbf{T}\mathbf{W}_{\mathbf{t}}^{1}))\dots),\\
\mathbf{Y} &= \boldsymbol{\psi}_{\mathbf{y}}(\mathbf{U}) = \mathbf{W}_{\mathbf{u}}^{L}\sigma(\dots\sigma(\mathbf{W}_{\mathbf{u}}^{2}\sigma(\mathbf{U}\mathbf{W}_{\mathbf{u}}^{1}))\dots).
\end{aligned}$$

Каждая нейросеть является суперпозицией последовательных умножений на матрицы параметров и применения поэлементных функций активаций.

Требуется найти такие параметры, при которых функция согласования g достигает своего максимума:

$$g(\boldsymbol{\tau},\boldsymbol{\nu}) \to \max_{\mathbf{W}}, \tag{15}$$

где $\mathbf{W} = \left\{\mathbf{W}_{\mathbf{x}}^{i}, \mathbf{W}_{\mathbf{y}}^{i}, \mathbf{W}_{\mathbf{t}}^{i}, \mathbf{W}_{\mathbf{u}}^{i}\right\}_{i=1}^{L}$.

Процесс согласования заключается в максимизации функции согласования $g(\boldsymbol{\tau},\boldsymbol{\nu})$ по параметрам нейросетей. В работе [?] рассматривается частный случай задачи (15). При использовании в качестве функции согласования корреляции между проекциями $g(\boldsymbol{\tau},\boldsymbol{\nu}) = \operatorname{corr}(\boldsymbol{\tau},\boldsymbol{\nu})$ частная производная функции согласования по первому аргументу принимает вид

$$\frac{\partial g(\boldsymbol{\tau},\boldsymbol{\nu})}{\partial \boldsymbol{\tau}} = \frac{1}{m-1}\left(\boldsymbol{\Sigma}_1^{-1/2}\mathbf{U}\mathbf{V}^{\intercal}\boldsymbol{\Sigma}_2^{-1/2}\mathbf{U} - \boldsymbol{\Sigma}_1^{-1/2}\mathbf{U}\mathbf{D}\mathbf{V}^{\intercal}\boldsymbol{\Sigma}_1^{-1/2}\right),$$

где $\mathbf{U},\mathbf{D},\mathbf{V} = \mathrm{SVD}(\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1/2}$, $\boldsymbol{\Sigma}_1 = \frac{1}{m-1}\mathbf{T}\mathbf{T}^{\intercal}$, $\boldsymbol{\Sigma}_2 = \frac{1}{m-1}\mathbf{U}\mathbf{U}^{\intercal}$, $\boldsymbol{\Sigma}_{12} = \frac{1}{m-1}\mathbf{T}\mathbf{U}^{\intercal}$. Аналогичное выражение имеет частная производная по второму аргументу. Полученное выражение для градиента позволяет построить эффективный алгоритм для решения задачи с использованием градиентных методов оптимизации.

Доказательство корректности алгоритма проекции в скрытое пространство.

Псевдокод метода регрессии PLS приведен в алгоритме 1. Алгоритм итеративно на каждом из l шагов вычисляет по одному столбцу $\boldsymbol{\tau}_k$, $\boldsymbol{\nu}_k$, $\mathbf{p}_k$, $\mathbf{q}_k$ матриц $\mathbf{T}$, $\mathbf{U}$, $\mathbf{P}$, $\mathbf{Q}$ соответственно. После вычисления следующего набора векторов из матриц $\mathbf{X}$, $\mathbf{Y}$ вычитаются очередные одноранговые аппроксимации. При этом предполагается, что исходные матрицы $\mathbf{X}$ и $\mathbf{Y}$ нормированы (имеют нулевое среднее и единичное среднее отклонение).

Вектора $\boldsymbol{\tau}_k$ и $\boldsymbol{\nu}_k$ из внутреннего цикла алгоритма 1 содержат информацию о матрице исходных объектов $\mathbf{X}$ и матрице целевых объектов $\mathbf{Y}$ соответственно. Блоки из шагов (6)–(7) и шагов (8)–(9) — аналоги метода PCA для матриц $\mathbf{X}$ и $\mathbf{Y}$ [?]. Последовательное выполнение блоков позволяет учесть взаимную связь между матрицами $\mathbf{X}$ и $\mathbf{Y}$.

Теоретическое обоснование метода PLS следует из следующих утверждений.

Algorithm 1 Алгоритм PLS

Вход: $\mathbf{X}, \mathbf{Y}, l$;
Выход: $\mathbf{T}, \mathbf{P}, \mathbf{Q}$;

1: нормировать матрицы $\mathbf{X}$ и $\mathbf{Y}$ по столбцам
2: инициализировать $\boldsymbol{\nu}_0$ (первый столбец матрицы $\mathbf{Y}$)
3: $\mathbf{X}_1 = \mathbf{X}; \mathbf{Y}_1 = \mathbf{Y}$
4: **для** $k = 1, \ldots, l$
5: **повторять**
6: $\mathbf{w}_k := \mathbf{X}_k^\intercal \boldsymbol{\nu}_{k-1} / (\boldsymbol{\nu}_{k-1}^\intercal \boldsymbol{\nu}_{k-1})$; $\mathbf{w}_k := \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}$
7: $\boldsymbol{\tau}_k := \mathbf{X}_k \mathbf{w}_k$
8: $\mathbf{c}_k := \mathbf{Y}_k^\intercal \boldsymbol{\tau}_k / (\boldsymbol{\tau}_k^\intercal \boldsymbol{\tau}_k)$; $\mathbf{c}_k := \frac{\mathbf{c}_k}{\|\mathbf{c}_k\|}$
9: $\boldsymbol{\nu}_k := \mathbf{Y}_k \mathbf{c}_k$
10: **пока** $\boldsymbol{\tau}_k$ не стабилизируется
11: $\mathbf{p}_k := \mathbf{X}_k^\intercal \boldsymbol{\tau}_k / (\boldsymbol{\tau}_k^\intercal \boldsymbol{\tau}_k)$, $\mathbf{q}_k := \mathbf{Y}_k^\intercal \boldsymbol{\tau}_k / (\boldsymbol{\tau}_k^\intercal \boldsymbol{\tau}_k)$
12: $\mathbf{X}_{k+1} := \mathbf{X}_k - \boldsymbol{\tau}_k \mathbf{p}_k^\intercal$
13: $\mathbf{Y}_{k+1} := \mathbf{Y}_k - \boldsymbol{\tau}_k \mathbf{q}_k^\intercal$

Утверждение 1. Максимизации ковариации между векторами $\boldsymbol{\tau}_k$ и $\boldsymbol{\nu}_k$ сохраняет дисперсию столбцов матриц $\mathbf{X}$ и $\mathbf{Y}$ и учитывает их линейную зависимость.

Доказательство. Утверждение следует из равенства

$$\operatorname{cov}(\boldsymbol{\tau}_k, \boldsymbol{\nu}_k) = \operatorname{corr}(\boldsymbol{\tau}_k, \boldsymbol{\nu}_k) \cdot \sqrt{\operatorname{var}(\boldsymbol{\tau}_k)} \cdot \sqrt{\operatorname{var}(\boldsymbol{\nu}_k)}.$$

Максимизация дисперсий векторов $\boldsymbol{\tau}_k$ и $\boldsymbol{\nu}_k$ отвечает за сохранение информации об исходных матрицах, корреляция между векторами отвечает взаимосвязи между $\mathbf{X}$ и $\mathbf{Y}$. □

Во внутреннем цикле алгоритма 1 вычисляются нормированные вектора весов $\mathbf{w}_k$ и $\mathbf{c}_k$. Из данных векторов строятся матрицы весов $\mathbf{W}$ и $\mathbf{C}$ соответственно.

Утверждение 2. В результате выполнения внутреннего цикла вектора $\mathbf{w}_k$ и $\mathbf{c}_k$ будут собственными векторами матриц $\mathbf{X}_k^\intercal \mathbf{Y}_k \mathbf{Y}_k^\intercal \mathbf{X}_k$ и $\mathbf{Y}_k^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal \mathbf{Y}_k$, соответствующими максимальным собственным значениям.

$$\mathbf{w}_k \propto \mathbf{X}_k^\intercal \boldsymbol{\nu}_{k-1} \propto \mathbf{X}_k^\intercal \mathbf{Y}_k \mathbf{c}_{k-1} \propto \mathbf{X}_k^\intercal \mathbf{Y}_k \mathbf{Y}_k^\intercal \boldsymbol{\tau}_{k-1} \propto \mathbf{X}_k^\intercal \mathbf{Y}_k \mathbf{Y}_k^\intercal \mathbf{X}_k \mathbf{w}_{k-1},$$
$$\mathbf{c}_k \propto \mathbf{Y}_k^\intercal \boldsymbol{\tau}_k \propto \mathbf{Y}_k^\intercal \mathbf{X}_k \mathbf{w}_k \propto \mathbf{Y}_k^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal \boldsymbol{\nu}_{k-1} \propto \mathbf{Y}_k^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal \mathbf{Y}_k \mathbf{c}_{k-1},$$

где символ $\propto$ означает равенство с точностью до мультипликативной константы.

Доказательство. Утверждение следует из того факта, что правила обновления векторов $\mathbf{w}_k$, $\mathbf{c}_k$ совпадают с итерацией алгоритма поиска максимального собственного значения. Данный алгоритм основан на следующем факте. Если матрица $\mathbf{A}$ диагонализуема, $\mathbf{x}$ — некоторый вектор, то

$$\lim_{k\to\infty} \mathbf{A}^k\mathbf{x} = \lambda_{\max}(\mathbf{A}) \cdot \mathbf{v}_{\max},$$

где $\lambda_{\max}(\mathbf{A})$ — максимальное собственное значение матрицы $\mathbf{A}$, $\mathbf{v}_{\max}$ — собственный вектор матрицы $\mathbf{A}$, соответствующий $\lambda_{\max}(\mathbf{A})$.

□

Утверждение 3. Обновление векторов по шагам (6)–(9) алгоритма 1 соответствует максимизации ковариации между векторами $\boldsymbol{\tau}_k$ и $\boldsymbol{\nu}_k$.

Доказательство. Максимальная ковариация между векторами $\boldsymbol{\tau}_k$ и $\boldsymbol{\nu}_k$ равна максимальному собственному значению матрицы $\mathbf{X}_k^\intercal\mathbf{Y}_k\mathbf{Y}_k^\intercal\mathbf{X}_k$:

$$\begin{aligned}\max_{\boldsymbol{\tau}_k,\boldsymbol{\nu}_k} \operatorname{cov}(\boldsymbol{\tau}_k,\boldsymbol{\nu}_k)^2 &= \max_{\substack{\|\mathbf{w}_k\|=1\\\|\mathbf{c}_k\|=1}} \operatorname{cov}\left(\mathbf{X}_k\mathbf{w}_k, \mathbf{Y}_k\mathbf{c}_k\right)^2 = \max_{\substack{\|\mathbf{w}_k\|=1\\\|\mathbf{c}_k\|=1}} \operatorname{cov}\left(\mathbf{c}_k^\intercal\mathbf{Y}_k^\intercal\mathbf{X}_k\mathbf{w}_k\right)^2 = \\ &= \max_{\|\mathbf{w}_k\|=1} \operatorname{cov}\left\|\mathbf{Y}_k^\intercal\mathbf{X}_k\mathbf{w}_k\right\|^2 = \max_{\|\mathbf{w}_k\|=1} \mathbf{w}_k^\intercal\mathbf{X}_k^\intercal\mathbf{Y}_k\mathbf{Y}_k^\intercal\mathbf{X}_k\mathbf{w}_k = \\ &= \lambda_{\max}\left(\mathbf{X}_k^\intercal\mathbf{Y}_k\mathbf{Y}_k^\intercal\mathbf{X}_k\right),\end{aligned}$$

где $\lambda_{\max}(\mathbf{A})$ — максимальное собственное значение матрицы $\mathbf{A}$. Применяя утверждение 2, получаем требуемое. □

После завершения внутреннего цикла на шаге (11) вычисляются вектора $\mathbf{p}_k$, $\mathbf{q}_k$ проецированием столбцов матриц $\mathbf{X}_k$ и $\mathbf{Y}_k$ на вектор $\boldsymbol{\tau}_k$. Для перехода на следующий шаг необходимо вычесть из матриц $\mathbf{X}_k$ и $\mathbf{Y}_k$ одноранговые аппроксимации $\boldsymbol{\tau}_k\mathbf{p}_k^\intercal$ и $\boldsymbol{\tau}_k\mathbf{q}_k^\intercal$

$$\begin{aligned}\mathbf{X}_{k+1} &= \mathbf{X}_k - \boldsymbol{\tau}_k\mathbf{p}_k^\intercal = \mathbf{X} - \sum_k \boldsymbol{\tau}_k\mathbf{p}_k^\intercal,\\ \mathbf{Y}_{k+1} &= \mathbf{Y}_k - \boldsymbol{\tau}_k\mathbf{q}_k^\intercal = \mathbf{Y} - \sum_k \boldsymbol{\tau}_k\mathbf{q}_k^\intercal.\end{aligned}$$

При этом каждый следующий вектор $\boldsymbol{\tau}_k$ оказывается ортогонален всем векторам $\boldsymbol{\tau}_j$, $j = 1,\ldots,k$.

Для получения прогнозов модели и нахождения параметров модели домножим справа формулу (8) на матрицу $\mathbf{W}$. Строки матрицы невязок $\mathbf{E}$ ортогональны столбцам матрицы $\mathbf{W}$, поэтому

$$\mathbf{XW} = \mathbf{TP}^\intercal\mathbf{W}.$$

Линейное преобразование между объектами в исходном и латентном пространстве имеет вид

$$\mathbf{T} = \mathbf{X}\mathbf{W}^*, \tag{16}$$

где $\mathbf{W}^* = \mathbf{W}(\mathbf{P}^\intercal\mathbf{W})^{-1}$.

Матрица параметров модели 4 находится из уравнений (9), (16)

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^\intercal + \mathbf{E} = \mathbf{X}\mathbf{W}^*\mathbf{Q}^\intercal + \mathbf{E} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}. \tag{17}$$

Таким образом, параметры модели (4) равны

$$\boldsymbol{\Theta} = \mathbf{W}(\mathbf{P}^\intercal\mathbf{W})^{-1}\mathbf{Q}^\intercal.$$

Финальная модель (17) является линейной, низкоразмерной в скрытом пространстве. Это снижает избыточность данных и повышает стабильность модели.

Аддитивная суперпозиция моделей декодирования.

Пусть $\mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\Theta}_1)$, $\mathbf{f}_2(\mathbf{x}_2, \boldsymbol{\Theta}_2)$ — линейные модели декодирования сигналов. Рассмотрим аддитивную суперпозицию моделей декодирования, то есть модель (2) вида

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \boldsymbol{\Theta}) + \boldsymbol{\varepsilon} = \boldsymbol{\Theta}_1\mathbf{x}_1 + \boldsymbol{\Theta}_2\mathbf{x}_2 + \boldsymbol{\varepsilon}, \tag{18}$$

где объект $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] \in \mathbb{R}^n$ состоит из двух подвекторов $\mathbf{x}_1 \in \mathbb{R}^k$, $\mathbf{x}_2 \in \mathbb{R}^{n-k}$. Тем самым матрица параметров $\boldsymbol{\Theta} \in \mathbb{R}^{n\times r}$ состоит из двух подматриц $\boldsymbol{\Theta}_1 \in \mathbb{R}^{k\times r}$, $\boldsymbol{\Theta}_2 \in \mathbb{R}^{n-k\times r}$.

Утверждение 4. Матрица параметров $\boldsymbol{\Theta}$ для модели (18), доставляющая минимум функции ошибки (5), имеет вид:

$$\boldsymbol{\Theta}_1 = (\mathbf{X}_1^\intercal\mathbf{M}_{\mathbf{X}_2}\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal\mathbf{M}_{\mathbf{X}_2}\mathbf{Y},$$
$$\boldsymbol{\Theta}_2 = (\mathbf{X}_2^\intercal\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\intercal\mathbf{M}_{\mathbf{X}_1}\mathbf{Y},$$

где $\mathbf{M}_{\mathbf{X}_1} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})$, $\mathbf{M}_{\mathbf{X}_2} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_2})$, $\mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_1(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal$, $\mathbf{P}_{\mathbf{X}_2} = \mathbf{X}_2(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1}\mathbf{X}_2^\intercal$.

Доказательство. Домножим уравнение $\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta}$ слева на матрицу $\mathbf{X}^\intercal$

$$\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\Theta} = \mathbf{X}^\intercal\mathbf{Y} \quad \Rightarrow \quad \begin{cases}\mathbf{X}_1^\intercal\mathbf{X}_1\boldsymbol{\Theta}_1 + \mathbf{X}_1^\intercal\mathbf{X}_2\boldsymbol{\Theta}_2 = \mathbf{X}_1^\intercal\mathbf{Y},\\ \mathbf{X}_2^\intercal\mathbf{X}_1\boldsymbol{\Theta}_1 + \mathbf{X}_2^\intercal\mathbf{X}_2\boldsymbol{\Theta}_2 = \mathbf{X}_2^\intercal\mathbf{Y}.\end{cases}$$

Выразим из этой системы параметры каждой отдельной модели в суперпозиции:

$$\begin{cases}\boldsymbol{\Theta}_1 = (\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal(\mathbf{Y} - \mathbf{X}_2\boldsymbol{\Theta}_2),\\ \boldsymbol{\Theta}_2 = (\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1}\mathbf{X}_2^\intercal(\mathbf{Y} - \mathbf{X}_1\boldsymbol{\Theta}_1).\end{cases}$$

Подставим полученные выражения для $\boldsymbol{\Theta}_1$ и $\boldsymbol{\Theta}_2$ в исходную систему:

$$\begin{cases}\mathbf{X}_1^\intercal\mathbf{P}_{\mathbf{X}_1}(\mathbf{Y} - \mathbf{X}_2\boldsymbol{\Theta}_2) + \mathbf{X}_1^\intercal\mathbf{X}_2\boldsymbol{\Theta}_2 = \mathbf{X}_1^\intercal\mathbf{Y},\\ \mathbf{X}_2^\intercal\mathbf{X}_1\boldsymbol{\Theta}_1 + \mathbf{X}_2^\intercal\mathbf{P}_{\mathbf{X}_2}(\mathbf{Y} - \mathbf{X}_1\boldsymbol{\Theta}_1) = \mathbf{X}_2^\intercal\mathbf{Y}.\end{cases}$$

Выразив матрицы параметров, получим требуемые выражения. □

Заметим, что матрицы $\mathbf{P}_{\mathbf{X}_1}$ и $\mathbf{P}_{\mathbf{X}_2}$ являются матрицами проекций на подпространства, образованные линейными оболочками столбцов матриц $\mathbf{X}_1$ и $\mathbf{X}_2$ соответственно. Таким образом матрицы $\mathbf{M}_{\mathbf{X}_1}$ и $\mathbf{M}_{\mathbf{X}_2}$ являются матрицами проекций на ортогональные подпространства.

Утверждение 5. Оптимальная подматрица $\mathbf{\Theta}_2$ в модели (18) является решением задачи регрессии

$$\|\mathbf{Y}_1 - \mathbf{X}_{21}\mathbf{\Theta}_2\| \to \min_{\mathbf{\Theta}_2}, \tag{19}$$

где $\mathbf{Y}_1 = \mathbf{M}_{\mathbf{X}_1}\mathbf{Y}$, $\mathbf{X}_{21} = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$.

Доказательство. Матрица $\mathbf{M}_{\mathbf{X}_1}$ является матрицей проекции на ортогональное подпространство, построенное на линейной оболочке столбцов матрицы $\mathbf{X}_1$. Таким образом, матрица $\mathbf{M}_{\mathbf{X}_1}$ является идемпотентной ($\mathbf{M}_{\mathbf{X}_1}{}^{\intercal}\mathbf{M}_{\mathbf{X}_1} = \mathbf{M}_{\mathbf{X}_1}$). Таким образом, используя утверждение 4,

$$\begin{aligned}\mathbf{\Theta}_2 = (\mathbf{X}_2^{\intercal}\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^{\intercal}\mathbf{M}_{\mathbf{X}_1}\mathbf{Y} = (\mathbf{X}_2^{\intercal}\mathbf{M}_{\mathbf{X}_1}^{\intercal}\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{-1}\mathbf{X}_2^{\intercal}\mathbf{M}_{\mathbf{X}_1}^{\intercal}\mathbf{M}_{\mathbf{X}_1}\mathbf{Y} = \\ = ((\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{\intercal}(\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{\intercal})^{-1}(\mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2)^{\intercal}(\mathbf{M}_{\mathbf{X}_1}\mathbf{Y}) = (\mathbf{X}_{21}^{\intercal}\mathbf{X}_{21})^{-1}\mathbf{X}_{21}^{\intercal}\mathbf{Y}_1.\end{aligned}$$

Согласно теореме Гаусса-Маркова $\mathbf{\Theta}_2$ является решением задачи регрессии (19). □

Таким образом для нахождении оптимальной модели декодирования $\mathbf{f}(\mathbf{x}_2, \mathbf{\Theta}_2)$ необходимо спроецировать матрицы $\mathbf{Y}$ и $\mathbf{X}_2$ на подпространства, ортогональные подпространству, образованному линейной оболочкой столбцов матрицы $\mathbf{X}_1$.

Похожие результаты были доказаны в эконометрике в работах [?, ?, ?]. Аналогичное утверждение верно и для матрицы $\mathbf{\Theta}_1$.

Утверждение 6. Если в задаче (18) $\operatorname{span}(\mathbf{X}_1) \cap \operatorname{span}(\mathbf{X}_2) = \varnothing$, то есть столбцы матрицы $\mathbf{X}_1$ ортогональны столбцам матрицы $\mathbf{X}_2$, то $\mathbf{\Theta}_2$ является решением задачи регрессии

$$\|\mathbf{Y} - \mathbf{X}_2\mathbf{\Theta}_2\| \to \min_{\mathbf{\Theta}_2}.$$

Доказательство. Используя факт, что $\mathbf{I} = \mathbf{P}_{\mathbf{X}_1} + \mathbf{M}_{\mathbf{X}_1}$ и $\mathbf{X}_2\mathbf{X}_1 = 0$, получаем

$$\begin{aligned}(\mathbf{X}_2^{\intercal}\mathbf{X}_2)^{-1}\mathbf{X}_2^{\intercal}\mathbf{Y} = (\mathbf{X}_2^{\intercal}\mathbf{X}_2)^{-1}\mathbf{X}_2^{\intercal}(\mathbf{P}_{\mathbf{X}_1} + \mathbf{M}_{\mathbf{X}_1})\mathbf{Y} = \\ = (\mathbf{X}_2^{\intercal}\mathbf{X}_2)^{-1}\mathbf{X}_2^{\intercal}\mathbf{M}_{\mathbf{X}_1}\mathbf{Y} = (\mathbf{X}_{21}^{\intercal}\mathbf{X}_{21})^{-1}\mathbf{X}_{21}^{\intercal}\mathbf{Y}_1.\end{aligned}$$

Используя утверждение 5, получаем требуемое утверждение. □

Данное утверждение показывает, что в случае независимых столбцов матриц $\mathbf{X}_1$ и $\mathbf{X}_2$ задача регрессии для аддитивной суперпозиции моделей (18) распадается на две независимые подзадачи.

Из выше доказанных утверждений прямо следует следующая теорема.

Теорема 1. Если в задаче (18) $\text{span}(\mathbf{X}_1) \neq \text{span}(\mathbf{X}_2)$, то ошибка аддитивной суперпозиции моделей не превышает ошибки каждой из отдельных моделей

$$\mathcal{L}(\mathbf{\Theta}^*, \mathbf{X}, \mathbf{Y}) \leqslant \mathcal{L}(\mathbf{\Theta}_1, \mathbf{X}_1, \mathbf{Y}),$$
$$\mathcal{L}(\mathbf{\Theta}^*, \mathbf{X}, \mathbf{Y}) \leqslant \mathcal{L}(\mathbf{\Theta}_2, \mathbf{X}_2, \mathbf{Y}).$$

Рассмотрим случай линейной авторегрессионной модели $\mathbf{f}_{\mathbf{y}}^{\text{AR}}$ из определения 4 и линейной регрессионной модели $\mathbf{f}_{\mathbf{xy}}^{\text{R}}$ из определения 5. Пусть модель декодирования $\mathbf{f}_{\mathbf{xy}} : \mathbb{R}^{h_x \times m} \times \mathbb{R}^{h_y \times r} \to \mathbb{R}^{p \times r}$ из определения 6 является аддитивной суперпозицией авторегрессионной и регрессионной моделей. Тогда при условии не совпадений подпространств, образованных линейными оболочками входами моделей, ошибка суперпозиции не будет превышать ошибок авторегрессионной и регрессионной моделей. Данное утверждение позволяет осуществлять выбор моделей в суперпозиции, основанный на анализе проекций подпространств, построенных на линейных оболочках исходных признаковых описаний.

Анализ линейных методов проекции в скрытое пространство.

Для проведения вычислительного эксперимента рассматриваются данные потребления электроэнергии. Временные ряды электроэнергии состоят из почасовых записей (52512 наблюдений). Строка матрицы $\mathbf{X}$ — локальная история сигнала за одну неделю $n = 24 \times 7$. Строка матрицы $\mathbf{Y}$ — локальный прогноз потребления электроэнергии в следующие 24 часа $r = 24$. В этом случае матрицы $\mathbf{X}$ и $\mathbf{Y}$ являются авторегрессионными матрицами.

Вычислительный эксперимент также проводился на данных электрокортикограмм (ECoG) из проекта NeuroTycho [**?**]. Данные ECoG состоят из 32-канальных сигналов напряжения, снятых с головного мозга. Цель состоит в предсказании по входному сигналу ECoG 3D позиции рук в последующие моменты времени. Исходные сигналы напряжения преобразуются в пространственно-временное представление с помощью вейвлет-преобразования с материнским вейвлетом Морле. Процедура извлечения признаков из исходных данных подробно описана в [**?**, **?**]. Описание исходного сигнала в каждый момент времени имеет размерность 32 (каналы) $\times$ 27 (частоты) = 864. Каждый объект представляет собой локальный отрезок времени длительностью $\Delta t = 1s$. Временной шаг между объектами $\delta t = 0.05s$. Матрицы имеют размеры $\mathbf{X} \in \mathbb{R}^{18900 \times 864}$ и $\mathbf{Y} \in \mathbb{R}^{18900 \times 3k}$, где k — число отсчётов времени прогнозирования. Данные разбиты на тренировочную и тестовую части в соотношении 0,67. Пример исходных сигналов мозга и соответствующей траектории руки показан на Рис. 2.

Введём среднеквадратичную ошибку для некоторых матриц $\mathbf{A} = [a_{ij}]$ и $\mathbf{B} = [b_{ij}]$

$$\text{MSE}(\mathbf{A}, \mathbf{B}) = \sum_{i,j} (a_{ij} - b_{ij})^2.$$

Для оценивания качества аппроксимации вычисляется значение нормированной

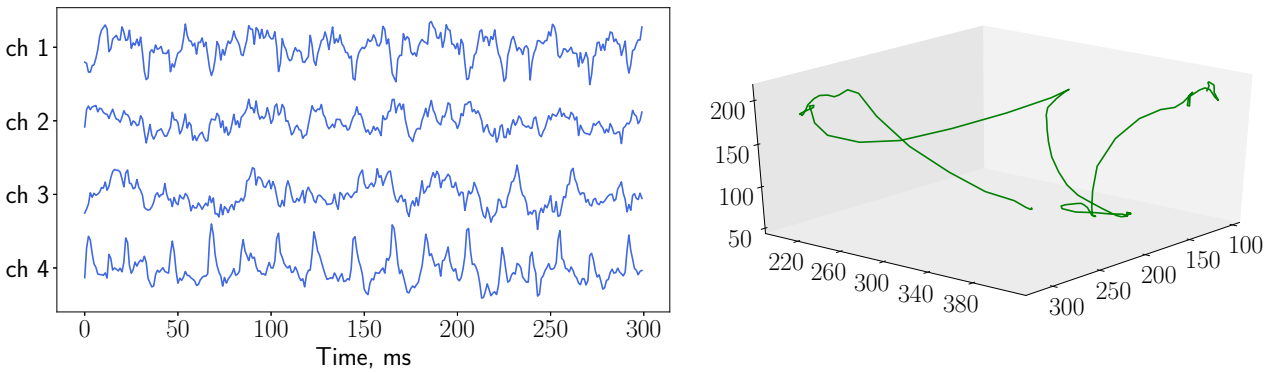

Рис. 2: Сигналы мозга (левый график) и 3D координаты руки (правый график)

среднеквадратичной ошибки

$$\text{NMSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{\text{MSE}(\mathbf{Y}, \hat{\mathbf{Y}})}{\text{MSE}(\mathbf{Y}, \bar{\mathbf{Y}})}, \tag{20}$$

где $\hat{\mathbf{Y}}$ — прогноз модели, $\bar{\mathbf{Y}}$ — константный прогноз средним значением по столбцам матрицы.

Результаты на данных электроэнергии. Для нахождения оптимальной размерности l латентного пространства все данные потребления электроэнергии были разбиты на обучающую и валидационную части. Обучающая выборка состоит из 700 объектов, валидационная из 370. Зависимость нормированной квадратичной ошибки (20) от размерности l латентного пространства представлена на Рис. **??**. Сначала ошибка резко падает при увеличении размерности скрытого пространства, а затем стабилизируется.

Минимальная ошибка наблюдается при $l = 14$. Построим прогноз потребления электроэнергии при данном l. Результат аппроксимации изображен на Рис. **??**. Алгоритм PLS восстановил авторегрессионную зависимость и обнаружил дневную сезонность.

Результаты на данных электрокортикограммы. На Рис. **??** представлена зависимость нормированной квадратичной ошибки (20) от размерности латентного пространства. Ошибка аппроксимации меняется незначительно при $l > 5$. Таким образом совместное описание пространственно-временного спектрального представления объектов и пространственного положения руки может быть представлено вектором размерности $l \ll n$. Зафиксируем $l = 5$. Пример аппроксимации положения руки изображен на Рис. **??**. Сплошными линиями изображены истинные координаты руки по всем осям, пунктирными линиями показана аппроксимация методом PLS.

0.1 Анализ нелинейных методов проекции в скрытое пространство

Цель вычислительного эксперимента — сравнительный анализ рассматриваемых моделей. Рассматриваются данные, для которых сложность класса

линейных методов неадекватно низка. Нелинейные модели позволяют получить точный прогноз при адекватной сложности.

Задача фильтрации шума. Проведем сравнение качества DeepCCA и CCA на задаче классификации зашумленных цифровых изображениях, представленных на Рис. **??**. Для этого используется набор данных MNIST [**?**], который состоит из 70 000 цифровых изображений 28×28 образцов рукописного написания цифр. Предлагается получить два новых набора данных $\mathbf{X}$ и $\mathbf{Y}$ следующим образом. Первый набор получается поворотом исходных изображений на угол в диапазоне $[\frac{-\pi}{4}, \frac{\pi}{4}]$. Для получения второго набора данных для каждой картинки из первого набора данных ставится в соответствие случайным образом картинка с той же цифрой, но с добавлением независимого случайного шума, распределенного равномерно на отрезке [0,1].

Таблица 2: Точность классификации линейного SVM для методов Deep CCA и CCA

Скользящий контроль	Deep CCA ($L = 3$)	CCA
Валидация	92,74%	76,21%
Тест	92,14%	76,07%

Применив к двум новым наборам данных DeepCCA или CCA, получаем новое низкоразмерное признаковое пространство, которое игнорирует шумы в исходных данных. Модель DeepCCA представляет собой нейронную сеть с $L = 3$ скрытыми слоями. Таким образом, получаем функции кодирования $\boldsymbol{\varphi}_{\mathbf{x}}$ и $\boldsymbol{\varphi}_{\mathbf{y}}$ для исходных наборов данных. На новых признаках, полученных разными моделями (DeepCCA и CCA), для первого набора данных, то есть на данных после применения функции кодирования $\boldsymbol{\varphi}_{\mathbf{x}}$ к первому набору исходных данных, обучим линейный SVM-классификатор. Показателем эффективности будет точность классификации линейного SVM на тестовых данных. В случае построения адекватного скрытого пространства полученные образы объектов будут линейно разделимы. Результаты эксперимента приведены в таблице 2. Точность классификации нелинейной модели существенно выше линейного метода CCA.

Задача восстановления изображений. Для анализа процедуры согласования проведен вычислительный эксперимент с предложенными нелинейными моделями. Для снижения размерности пространства используются нейросетевые модели автоэнкодера с согласованием скрытого пространства (15). В качестве базовых моделей используются модель автоэнкодера без согласования скрытых пространств, а также линейный PLS. В качестве исходного набора данных используется набор данных MNIST [**?**]. Каждое изображение поделено на левую и правую части, как показано на Рис. **??**. Модель по левому изображению восстанавливает правое изображение.

Модель EncNet1 — нейронная сеть с нелинейными функциями активации, которая обучается на данных после преобразования их автоэнкодером. Модель

LinNet1 — нейронная сеть с одним линейным слоем, которая также обучается на преобразованных данных. Для EncNet1 и LinNet1 автоэнкодеры для исходных и целевых объектов используют совместную функцию потерь, которая связывает выходы энкодеров. Модели EncNet2 и LinNet2 устроены аналогично EncNet1 и LinNet1 соответственно, но в автоэнкодерах нет совместной функции потерь. Модель DumbNet — нейронная сеть, которая обучается на исходных данных и имеет такую же структуру, что и EncNet, то есть имеет такое же число слоев и в каждом слое такое же количество нейронов, что и у EncNet.

Для оценки качества моделей вычислялась среднеквадратичная ошибка. Примеры восстановленных изображений показаны на Рис. **??**. Качество моделей, а также их сложность представлены в таблице 3. На Рис. **??** показано, что предложенные модели EncNet и LinNet позволяют получить более четкие и различимые изображения, в отличие от базовой нелинейной модели DumbNet и линейной модели PLS. Несмотря на заметное улучшение визуального качества изображений, ошибка предложенных моделей выше, чем у модели DumbNet. Это связано с тем, что среднеквадратичная ошибка оказалась неадекватной метрикой в пространстве изображений. Нахождение оптимальной метрики для оценки качества предложенных методов может быть одним из возможных направлений развития текущего эксперимента.

Таблица 3: Квадратичная ошибка для нелинейных моделей в задаче восстановления правой части изображения по левой

	EncNet1	LinNet1	EncNet2	LinNet2	DumbNet	PLS
Число параметров, тыс.	283	239	283	239	283	–
Ошибка на тесте	0,147	0,235	0,149	0,236	0,128	0,188

Глава 3. Выбор признаков в задаче декодирования сигналов.

Задача выбора признаков заключается в поиске оптимального подмножества $\mathcal{A} \subset \{1, \ldots, n\}$ индексов признаков среди всех возможных $2^n - 1$ вариантов. Существует взаимооднозначное отображение между подмножеством $\mathcal{A}$ и булевым вектором $\mathbf{a} \in \{0, 1\}^n$, компоненты которого указывают, выбран ли признак. Для нахождения оптимального вектора $\mathbf{a}$ введем функцию ошибки выбора признаков $S(\mathbf{a}, \mathbf{X}, \mathbf{Y})$. Проблема выбора признаков принимает вид:

$$\mathbf{a} = \arg\min_{\mathbf{a}' \in \{0,1\}^n} S(\mathbf{a}', \mathbf{X}, \mathbf{Y}). \tag{21}$$

Целью выбора признаков является построение функции $S(\mathbf{a}, \mathbf{X}, \mathbf{Y})$. Конкретные примеры данной функции для рассматриваемых методов выбора признаков приведены ниже и обобщены в таблице 4.

Задача (21) имеет дискретную область определения $\{0, 1\}^n$. Для решения данной задачи применяется релаксация задачи (21) к непрерывной области определения $[0, 1]^n$. Релаксированная задача выбора признаков имеет следующий

вид:

$$\mathbf{z} = \arg\min_{\mathbf{z}' \in [0,1]^n} S(\mathbf{z}', \mathbf{X}, \mathbf{Y}). \tag{22}$$

Здесь компоненты вектора $\mathbf{z}$ — значения нормированных коэффициентов значимости признаков. Сначала решается задача (22), для получения вектора значимостей $\mathbf{z}$. Затем решение (21) восстанавливается с помощью отсечения по порогу следующим образом:

$$\mathbf{a} = [a_j]_{j=1}^n, \quad a_j = \begin{cases} 1, & z_j > \tau; \\ 0, & \text{в противном случае.} \end{cases} \tag{23}$$

τ — гиперпараметр, который может быть подобран вручную или выбран с помощью кросс-валидации.

Как только решение $\mathbf{a}$ задачи (21) получено, задача (5) принимает вид:

$$\mathcal{L}(\boldsymbol{\Theta}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\Theta}_{\mathcal{A}}\|_2^2 \to \min_{\boldsymbol{\Theta}_{\mathcal{A}}},$$

где индекс $\mathcal{A}$ обозначает подматрицу со столбцами, индексы которых содержатся в $\mathcal{A}$.

Выбор признаков с помощью квадратичного программирования.

Если между столбцами матрицы исходных объектов $\mathbf{X}$ существует линейная зависимость, то решение задачи линейной регрессии

$$\|\boldsymbol{\upsilon} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \to \min_{\boldsymbol{\theta} \in \mathbb{R}^n}. \tag{24}$$

оказывается неустойчивым. Методы выбора признаков находят подмножество $\mathcal{A} \in \{1, \ldots, n\}$ оптимальных столбцов матрицы $\mathbf{X}$.

Метод QPFS [**?**] выбирает некоррелированные признаки, релевантные целевому вектору $\boldsymbol{\upsilon}$. Чтобы формализовать этот подход, введем две функции: $\mathrm{Sim}(\mathbf{X})$ и $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\upsilon})$. $\mathrm{Sim}(\mathbf{X})$ контролирует избыточность между признаками, $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\upsilon})$ содержит релевантности между каждым признаком и целевым вектором. Мы хотим минимизировать функцию Sim и максимизировать Rel одновременно.

QPFS предлагает явный способ построения функций Sim и Rel. Метод минимизирует следующую функцию ошибки

$$\underbrace{\mathbf{z}^\intercal \mathbf{Q} \mathbf{z}}_{\mathrm{Sim}} - \alpha \cdot \underbrace{\mathbf{b}^\intercal \mathbf{z}}_{\mathrm{Rel}} \to \min_{\substack{\mathbf{z} \in \mathbb{R}_+^n \\ \|\mathbf{z}\|_1 = 1}}. \tag{25}$$

Элементы матрицы парных взаимодействий $\mathbf{Q} \in \mathbb{R}^{n \times n}$ содержат коэффициенты попарного сходства между признаками. Вектор релевантностей признаков $\mathbf{b} \in \mathbb{R}^n$ выражает сходство между каждым признаком и целевым вектором $\boldsymbol{\upsilon}$. Нормированный вектор $\mathbf{z}$ отражает значимость каждого признака. Функция ошибки (25) штрафует зависимые признаки функцией Sim и штрафует признаки,

не релевантные к целевой переменной функцией Rel. Параметр α позволяет контролировать компромисс между Sim и Rel. Авторы оригинальной статьи QPFS [**?**] предложили способ выбора α, чтобы уравновесить вклад членов $\mathrm{Sim}(\mathbf{X})$ и $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\upsilon})$

$$\alpha = \frac{\overline{\mathbf{Q}}}{\overline{\mathbf{Q}} + \overline{\mathbf{b}}}, \quad \text{где } \overline{\mathbf{Q}} = \mathrm{mean}(\mathbf{Q}), \;\; \overline{\mathbf{b}} = \mathrm{mean}(\mathbf{b}).$$

Чтобы выделить оптимальное подмножество признаков, применяется отсечение по порогу (23).

Для измерения сходства используется выборочный коэффициент корреляции Пирсона между парами признаков для функции Sim, и между признаками и целевым вектором для функции Rel:

$$\mathbf{Q} = \left[|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)|\right]_{i,j=1}^{n}, \quad \mathbf{b} = \left[|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\upsilon})|\right]_{i=1}^{n}. \tag{26}$$

Здесь

$$\mathrm{corr}(\boldsymbol{\chi}, \boldsymbol{\upsilon}) = \frac{\sum_{i=1}^{m}(\boldsymbol{\chi}_i - \overline{\boldsymbol{\chi}})(\boldsymbol{\upsilon}_i - \overline{\boldsymbol{\upsilon}})}{\sqrt{\sum_{i=1}^{m}(\boldsymbol{\chi}_i - \overline{\boldsymbol{\chi}})^2 \sum_{i=1}^{m}(\boldsymbol{\upsilon}_i - \overline{\boldsymbol{\upsilon}})^2}}.$$

Другие способы определения $\mathbf{Q}$ и $\mathbf{b}$ рассматриваются в [**?**]. В работе [**?**] показано, что метод QPFS превосходит многие существующие методы выбора признаков на различных внешних критериях качества.

Задача (25) является выпуклой, если матрица $\mathbf{Q}$ является неотрицательно определенной. В общем случае это не всегда верно. Чтобы удовлетворить этому условию спектр матрицы $\mathbf{Q}$ смещается, и матрица $\mathbf{Q}$ заменяется на $\mathbf{Q} - \lambda_{\min}\mathbf{I}$, где $\lambda_{\min}$ является минимальным собственным значением $\mathbf{Q}$.

Методы выбора признаков для случая векторной целевой переменной.

В данном разделе описаны предлагаемые методы выбора признаков для случая векторной целевой переменной. В этом случае компоненты целевой переменной могут коррелировать между собой. Предлагаются методы, учитывающие зависимости как в исходном, так и в целевом пространствах.

Агрегация релевантностей целевых векторов. В работе [**?**], чтобы применить метод QPFS к векторному случаю ($r > 1$), релевантности признаков агрегируются по всем r компонентам целевой переменной. Член $\mathrm{Sim}(\mathbf{X})$ остаётся без изменений, матрица парных взаимодействий $\mathbf{Q}$ определяется как (26). Вектор релевантностей $\mathbf{b}$ агрегируется по всем компонентам целевой переменной и определяется как

$$\mathbf{b} = \left[\sum_{k=1}^{r} |\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\upsilon}_k)|\right]_{i=1}^{n}.$$

Недостатком такого подхода является отсутствие учёта зависимостей в столбцах

матрицы $\mathbf{Y}$. Рассмотрим следующий пример:

$$\mathbf{X} = [\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3], \quad \mathbf{Y} = [\underbrace{\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_1}_{r-1}, \boldsymbol{\upsilon}_2].$$

Пусть матрица $\mathbf{X}$ содержит 3 столбца, матрица $\mathbf{Y}$ — r столбцов, где первые $r-1$ компонент целевой переменной идентичны. Попарные сходства признаков задаются матрицей $\mathbf{Q}$. Матрица $\mathbf{B}$ содержит попарные сходства признаков и целевых столбцов. Вектор $\mathbf{b}$ получен суммированием матрицы $\mathbf{B}$ по столбцами

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.4 & \ldots & 0.4 & 0 \\ 0.5 & \ldots & 0.5 & 0.8 \\ 0.8 & \ldots & 0.8 & 0.1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (r-1) \cdot 0.4 + 0 \\ (r-1) \cdot 0.5 + 0.8 \\ (r-1) \cdot 0.8 + 0.1 \end{bmatrix}. \tag{27}$$

Пусть необходимо выбрать только 2 признака. В данном случае оптимальным подмножеством признаков является $[\boldsymbol{\chi}_1, \boldsymbol{\chi}_2]$. Признак $\boldsymbol{\chi}_2$ предсказывает второй целевой столбец $\boldsymbol{\upsilon}_2$, комбинация признаков $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ прогнозирует первый целевой столбец $\boldsymbol{\upsilon}_1$. Метод QPFS для $r = 2$ дает решение $\mathbf{z} = [0.37, 0.61, 0.02]$. Это совпадает с описанным решением. Однако, если добавить коллинеарные столбцы в матрицу $\mathbf{Y}$ и увеличить r до 5, то решением QPFS будет $\mathbf{z} = [0.40, 0.17, 0.43]$. Здесь потерян признак $\boldsymbol{\chi}_2$ и выбран избыточный признак $\boldsymbol{\chi}_3$. В следующих подразделах предлагаются обобщения метода QPFS, которые позволяют бороться с проблемой данного примера.

Симметричный учёт значимости признаков и целевых переменных. Чтобы учесть зависимости в столбцах матрицы $\mathbf{Y}$, обобщим функцию QPFS (25) для случая векторной целевой переменной ($r > 1$). Добавим член $\mathrm{Sim}(\mathbf{Y})$ и изменим член $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ следующим образом:

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^\top \mathbf{Q}_x \mathbf{z}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{z}_x^\top \mathbf{B} \mathbf{z}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{z}_y^\top \mathbf{Q}_y \mathbf{z}_y}_{\mathrm{Sim}(\mathbf{Y})} \to \min_{\substack{\mathbf{z}_x \geqslant \mathbf{0}_n, \mathbf{1}_n^\top \mathbf{z}_x = 1 \\ \mathbf{z}_y \geqslant \mathbf{0}_r, \mathbf{1}_r^\top \mathbf{z}_y = 1}}. \tag{28}$$

Определим элементы матриц $\mathbf{Q}_x \in \mathbb{R}^{n \times n}$, $\mathbf{Q}_y \in \mathbb{R}^{r \times r}$ и $\mathbf{B} \in \mathbb{R}^{n \times r}$ следующим образом:

$$\mathbf{Q}_x = \left[|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)| \right]_{i,j=1}^{n}, \quad \mathbf{Q}_y = \left[|\mathrm{corr}(\boldsymbol{\upsilon}_i, \boldsymbol{\upsilon}_j)| \right]_{i,j=1}^{r}, \quad \mathbf{B} = \left[|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\upsilon}_j)| \right]_{\substack{i=1,\ldots,n \\ j=1,\ldots,r}}.$$

Вектор $\mathbf{z}_x$ содержит коэффициенты значимости признаков, $\mathbf{z}_y$ — коэффициенты значимости целевых векторов. Коррелированные целевые столбцы штрафуются членом $\mathrm{Sim}(\mathbf{Y})$ и получают более низкие значения значимости.

Коэффициенты α_1, α_2, и α_3 контролируют влияние каждого члена на функцию (28) и удовлетворяют следующим условиям:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1, \quad \alpha_i \geqslant 0,\ i = 1, 2, 3.$$

Утверждение 7. Баланс между $\mathrm{Sim}(\mathbf{X})$, $\mathrm{Rel}(\mathbf{X},\mathbf{Y})$ и $\mathrm{Sim}(\mathbf{Y})$ в задаче (28) достигается при:

$$\alpha_1 \propto \overline{\mathbf{Q}}_y\overline{\mathbf{B}}, \quad \alpha_2 \propto \overline{\mathbf{Q}}_x\overline{\mathbf{Q}}_y, \quad \alpha_3 \propto \overline{\mathbf{Q}}_x\overline{\mathbf{B}}. \tag{29}$$

Доказательство. Значения α_1, α_2, и α_3 получаются путем решения следующих уравнений:

$$\begin{aligned}&\alpha_1 + \alpha_2 + \alpha_3 = 1,\\ &\alpha_1\overline{\mathbf{Q}}_x = \alpha_2\overline{\mathbf{B}} = \alpha_3\overline{\mathbf{Q}}_y.\end{aligned}$$

Здесь $\overline{\mathbf{Q}}_x$, $\overline{\mathbf{B}}$ и $\overline{\mathbf{Q}}_y$ — средние значения соответствующих матриц $\mathbf{Q}_x$, $\mathbf{B}$ и $\mathbf{Q}_y$ членов $\mathrm{Sim}(\mathbf{X})$, $\mathrm{Rel}(\mathbf{X},\mathbf{Y})$ и $\mathrm{Sim}(\mathbf{Y})$. □

Для изучения зависимости $\mathrm{Sim}(\mathbf{Y})$ на функцию (28), зафиксируем соотношение между α_1 и α_2:

$$\alpha_1 = \frac{(1-\alpha_3)\overline{\mathbf{B}}}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}, \quad \alpha_2 = \frac{(1-\alpha_3)\overline{\mathbf{Q}}_x}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}, \quad \alpha_3 \in [0,1]. \tag{30}$$

Применим предложенный метод к приведенному примеру (27). Матрица $\mathbf{Q}$ соответствует матрице $\mathbf{Q}_x$. Определим матрицы $\mathbf{Q}_y$ как $\mathrm{corr}(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2) = 0.2$, а все остальные элементы зададим 1. Рисунок **??** показывает значение векторов значимостей признаков $\mathbf{z}_x$ и целевых векторов $\mathbf{z}_y$ в зависимости от значения коэффициента α_3. Если α_3 мало, значимости всех целевых векторов не различимы и значимость признака $\boldsymbol{\chi}_3$ выше значимости признака $\boldsymbol{\chi}_2$. При увеличении α_3 до 0.2, коэффициент значимости $\mathbf{z}_{y,5}$ целевого вектора $\boldsymbol{\upsilon}_5$ увеличивается наряду со значимостью признака $\boldsymbol{\chi}_2$.

Минимаксная постановка задачи выбора признаков. Функция (28) является симметричной по отношению к $\mathbf{z}_x$ и $\mathbf{z}_y$. Она штрафует признаки, которые коррелированы и не имеют отношения к целевым векторам. Кроме того, она штрафует целевые вектора, которые коррелированы между собой и недостаточно коррелируют с признаками. Это приводит к малым значениям значимостей для целевых векторов, которые слабо коррелируют с признаками, и большим значениям для целевых векторов, которые сильно коррелируют с признаками. Этот результат противоречит интуиции. Цель — предсказать все целевые вектора, особенно те, которые слабо коррелируют с признаками. Сформулируем две взаимосвязанные задачи:

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}}\mathbf{Q}_x\mathbf{z}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} \to \min_{\substack{\mathbf{z}_x \geqslant \mathbf{0}_n,\\ \mathbf{1}_n^{\mathsf{T}}\mathbf{z}_x = 1}}, \tag{31}$$

$$\alpha_3 \cdot \underbrace{\mathbf{z}_y^{\mathsf{T}}\mathbf{Q}_y\mathbf{z}_y}_{\mathrm{Sim}(\mathbf{Y})} + \alpha_2 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} \to \min_{\substack{\mathbf{z}_y \geqslant \mathbf{0}_r,\\ \mathbf{1}_r^{\mathsf{T}}\mathbf{z}_y = 1}}. \tag{32}$$

Разница между (31) и (32) заключается в знаке перед членом Rel. В пространстве исходных объектов нерелевантные признаки должны иметь меньшие значения значимости. В то же время целевые вектора, не релевантные признакам, должны иметь большую значимость. Задачи (31) и (32) объединяются в совместную минимакс или максмин постановку

$$\min_{\substack{\mathbf{z}_x \geqslant \mathbf{0}_n \\ \mathbf{1}_n^\intercal \mathbf{z}_x = 1}} \max_{\substack{\mathbf{z}_y \geqslant \mathbf{0}_r \\ \mathbf{1}_r^\intercal \mathbf{z}_y = 1}} f(\mathbf{z}_x, \mathbf{z}_y), \quad \left(\text{или} \max_{\substack{\mathbf{z}_y \geqslant \mathbf{0}_r \\ \mathbf{1}_r^\intercal \mathbf{z}_y = 1}} \min_{\substack{\mathbf{z}_x \geqslant \mathbf{0}_n \\ \mathbf{1}_n^\intercal \mathbf{z}_x = 1}} f(\mathbf{z}_x, \mathbf{z}_y) \right), \tag{33}$$

где

$$f(\mathbf{z}_x, \mathbf{z}_y) = \alpha_1 \cdot \underbrace{\mathbf{z}_x^\intercal \mathbf{Q}_x \mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{z}_x^\intercal \mathbf{B} \mathbf{z}_y}_{\text{Rel}(\mathbf{X},\mathbf{Y})} - \alpha_3 \cdot \underbrace{\mathbf{z}_y^\intercal \mathbf{Q}_y \mathbf{z}_y}_{\text{Sim}(\mathbf{Y})}.$$

Теорема 2. Для положительно определенной матрицы $\mathbf{Q}_x$ и $\mathbf{Q}_y$, максмин и минимакс задачи (33) имеют одинаковое оптимальное значение.

Доказательство. Введём обозначения

$$\mathbb{C}^n = \{\mathbf{z} : \mathbf{z} \geqslant \mathbf{0}_n,\ \mathbf{1}_n^\intercal \mathbf{z} = 1\}, \quad \mathbb{C}^r = \{\mathbf{z} : \mathbf{z} \geqslant \mathbf{0}_r,\ \mathbf{1}_r^\intercal \mathbf{z} = 1\}.$$

Множества $\mathbb{C}^n$ и $\mathbb{C}^r$ компактные и выпуклые. Функция $f : \mathbb{C}^n \times \mathbb{C}^r \to \mathbb{R}$ является непрерывной. Если $\mathbf{Q}_x$ и $\mathbf{Q}_y$ положительно определенны, функция f является выпукло-вогнутой. Таким образом $f(\cdot, \mathbf{z}_y) : \mathbb{C}^n \to \mathbb{R}$ выпуклая при фиксированном $\mathbf{z}_y$, а $f(\mathbf{z}_x, \cdot) : \mathbb{C}^r \to \mathbb{R}$ вогнута при фиксированном $\mathbf{z}_x$. В этом случае по теореме Неймана о минимаксе

$$\min_{\mathbf{z}_x \in \mathbb{C}^n} \max_{\mathbf{z}_y \in \mathbb{C}^r} f(\mathbf{z}_x, \mathbf{z}_y) = \max_{\mathbf{z}_y \in \mathbb{C}^r} \min_{\mathbf{z}_x \in \mathbb{C}^n} f(\mathbf{z}_x, \mathbf{z}_y).$$

□

Для решения минимакс задачи (33), зафиксируем некоторый $\mathbf{z}_x \in \mathbb{C}^n$. Для фиксированного вектора $\mathbf{z}_x$ решаем задачу

$$\max_{\mathbf{z}_y \in \mathbb{C}_r} f(\mathbf{z}_x, \mathbf{z}_y) = \max_{\substack{\mathbf{z}_y \geqslant \mathbf{0}_r \\ \mathbf{1}_r^\intercal \mathbf{z}_y = 1}} \left[\alpha_1 \cdot \mathbf{z}_x^\intercal \mathbf{Q}_x \mathbf{z}_x - \alpha_2 \cdot \mathbf{z}_x^\intercal \mathbf{B} \mathbf{z}_y - \alpha_3 \cdot \mathbf{z}_y^\intercal \mathbf{Q}_y \mathbf{z}_y \right]. \tag{34}$$

Лагранжиан для данной задачи:

$$L(\mathbf{z}_x, \mathbf{z}_y, \lambda, \boldsymbol{\mu}) = \alpha_1 \cdot \mathbf{z}_x^\intercal \mathbf{Q}_x \mathbf{z}_x - \alpha_2 \cdot \mathbf{z}_x^\intercal \mathbf{B} \mathbf{z}_y - \alpha_3 \cdot \mathbf{z}_y^\intercal \mathbf{Q}_y \mathbf{z}_y + \lambda \cdot (\mathbf{1}_r^\intercal \mathbf{z}_y - 1) + \boldsymbol{\mu}^\intercal \mathbf{z}_y.$$

Здесь вектор множителей Лагранжа $\boldsymbol{\mu}$, который соответствует ограничениям на неравенства $\mathbf{z}_y \geqslant \mathbf{0}_r$, является неотрицательным. Двойственной задачей является

$$\min_{\lambda,\, \boldsymbol{\mu} \geqslant \mathbf{0}_r} g(\mathbf{z}_x, \lambda, \boldsymbol{\mu}) = \min_{\lambda,\, \boldsymbol{\mu} \geqslant \mathbf{0}_r} \left[\max_{\mathbf{z}_y \in \mathbb{R}^r} L(\mathbf{z}_x, \mathbf{z}_y, \lambda, \boldsymbol{\mu}) \right]. \tag{35}$$

Для задачи квадратичного программирования (34) с положительно определенными матрицами $\mathbf{Q}_x$ и $\mathbf{Q}_y$ выполняются условия сильной двойственности. Таким образом, оптимальное значение (34) равно оптимальному значению (35). Это позволяет перейти от решения задачи (33) к решению задачи

$$\min_{\mathbf{z}_x \in \mathbb{C}^n,\, \lambda,\, \boldsymbol{\mu} \geqslant \mathbf{0}_r} g(\mathbf{z}_y, \lambda, \boldsymbol{\mu}). \tag{36}$$

Полагая градиент $\nabla_{\mathbf{z}_y} L(\mathbf{z}_x, \mathbf{z}_y, \lambda, \boldsymbol{\mu})$ равным нулю, получим оптимальное значение $\mathbf{z}_y$:

$$\mathbf{z}_y = \frac{1}{2\alpha_3} \mathbf{Q}_y^{-1} \left(-\alpha_2 \cdot \mathbf{B}^\intercal \mathbf{z}_x + \lambda \cdot \mathbf{1}_r + \boldsymbol{\mu} \right).$$

Двойственная функция принимает вид

$$\begin{aligned} g(\mathbf{z}_x, \lambda, \boldsymbol{\mu}) = \max_{\mathbf{z}_y \in \mathbb{R}^r} L(\mathbf{z}_x, \mathbf{z}_y, \lambda, \boldsymbol{\mu}) = \mathbf{z}_x^\intercal \left(-\frac{\alpha_2^2}{4\alpha_3} \cdot \mathbf{B}\mathbf{Q}_y^{-1}\mathbf{B}^\intercal - \alpha_1 \cdot \mathbf{Q}_x \right) \mathbf{z}_x \\ - \frac{1}{4\alpha_3}\lambda^2 \cdot \mathbf{1}_r^\intercal \mathbf{Q}_y^{-1} \mathbf{1}_r - \frac{1}{4\alpha_3} \cdot \boldsymbol{\mu}^\intercal \mathbf{Q}_y^{-1} \boldsymbol{\mu} + \frac{\alpha_2}{2\alpha_3} \lambda \cdot \mathbf{1}_r^\intercal \mathbf{Q}_y^{-1} \mathbf{B}^\intercal \mathbf{z}_x \\ - \frac{1}{2\alpha_3} \lambda \cdot \mathbf{1}_r^\intercal \mathbf{Q}_y^{-1} \boldsymbol{\mu} + \frac{\alpha_2}{2\alpha_3} \cdot \boldsymbol{\mu}^\intercal \mathbf{Q}_y^{-1} \mathbf{B}^\intercal \mathbf{z}_x + \lambda. \end{aligned} \tag{37}$$

Тем самым задача (36) является квадратичной задачей с $n + r + 1$ переменными.

Несимметричный учёт значимостей признаков и целевых переменных. Естественным способом преодоления проблемы метода SymImp является добавление штрафа для целевых векторов, которые коррелируют с признаками. Добавим линейный член $\mathbf{b}^\intercal \mathbf{z}_y$ в член $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ следующим образом:

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^\intercal \mathbf{Q}_x \mathbf{z}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\left(\mathbf{z}_x^\intercal \mathbf{B} \mathbf{z}_y - \mathbf{b}^\intercal \mathbf{z}_y \right)}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{z}_y^\intercal \mathbf{Q}_y \mathbf{z}_y}_{\mathrm{Sim}(\mathbf{Y})} \to \min_{\substack{\mathbf{z}_x \geqslant \mathbf{0}_n,\, \mathbf{1}_n^\intercal \mathbf{z}_x = 1 \\ \mathbf{z}_y \geqslant \mathbf{0}_r,\, \mathbf{1}_r^\intercal \mathbf{z}_y = 1}}. \tag{38}$$

Утверждение 8. Пусть вектор $\mathbf{b}$ равен

$$b_j = \max_{i=1,\dots n} [\mathbf{B}]_{i,j}.$$

Тогда значение коэффициентов значимостей вектора $\mathbf{z}_y$ будут неотрицательными в $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ для задачи (38).

Доказательство. Утверждение следует из факта

$$\sum_{i=1}^n z_i b_{ij} \leqslant \left(\sum_{i=1}^n z_i \right) \max_{i=1,\dots n} b_{ij} = \max_{i=1,\dots n} b_{ij},$$

где $z_i \geqslant 0$ и $\sum_{i=1}^n z_i = 1$. □

Следовательно, функция (38) штрафует в меньшей мере признаки, которые имеют отношение к целевым векторам, и целевые вектора, которые недостаточно коррелированы с признаками.

Утверждение 9. Баланс между членами $\mathrm{Sim}(\mathbf{X})$, $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ и $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ для задачи (38) достигается при следующих коэффициентах:

$$\alpha_1 \propto \overline{\mathbf{Q}}_y \left(\overline{\mathbf{b}} - \overline{\mathbf{B}}\right), \quad \alpha_2 \propto \overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y, \quad \alpha_3 \propto \overline{\mathbf{Q}}_x \overline{\mathbf{B}}.$$

Доказательство. Необходимые значения α_1, α_2, и α_3 являются решением следующей системы уравнений:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1, \tag{39}$$

$$\alpha_1 \overline{\mathbf{Q}}_x = \alpha_2 \overline{\mathbf{B}}, \tag{40}$$

$$\alpha_2 \left(\overline{\mathbf{b}} - \overline{\mathbf{B}}\right) = \alpha_3 \overline{\mathbf{Q}}_y. \tag{41}$$

Здесь, в (40) уравновешены $\mathrm{Sim}(\mathbf{X})$ с первым слагаемым $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$, а в (41) уравновешены $\mathrm{Sim}(\mathbf{Y})$ с $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$. □

Утверждение 10. Для случая $r = 1$, предложенные функции (28), (33) и (38) совпадают с оригинальным методом QPFS (25).

Доказательство. Если r равно 1, то $\mathbf{Q}_y = q_y$ — скаляр, $\mathbf{z}_y = 1$ и $\mathbf{B} = \mathbf{b}$. Задачи (28), (33) и (38) принимают вид

$$\alpha_1 \cdot \mathbf{z}_x^\intercal \mathbf{Q}_x \mathbf{z}_x - \alpha_2 \cdot \mathbf{z}_x^\intercal \mathbf{b} \to \min_{\mathbf{z}_x \geqslant \mathbf{0}_n,\, \mathbf{1}_n^\intercal \mathbf{z}_x = 1}.$$

При $\alpha = \frac{\alpha_2}{\alpha_1 + \alpha_2}$ последняя задачи принимает вид (25). □

Таблица 4 демонстрирует основные идеи и функции ошибок для каждого метода. RelAgg является базовой стратегией и не учитывает корреляции в целевом пространстве. SymImp штрафует попарные корреляции между целевыми векторами. MinMax более чувствителен к целевым векторам, которые трудно предсказать. Стратегия Asymimp добавляет линейный член к функции SymImp, чтобы сделать вклад признаков и целевых векторов асимметричным.

Анализ методов учета значимостей целевых переменных.

Внешние критерии качества. Для оценки предложенных методов выбора признаков, введём критерии оценки качества выбранного подмножества признаков. Определим коэффициент мультикорреляции как среднее значение коэффициента множественной корреляции следующим образом:

$$R^2 = \frac{1}{r} \mathrm{tr} \left(\mathbf{C}^\intercal \mathbf{R}^{-1} \mathbf{C}\right), \quad \text{где } \mathbf{C} = [\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\upsilon}_j)]_{\substack{i=1,\ldots,n \\ j=1,\ldots,r}},\ \mathbf{R} = [\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)]_{i,j=1}^{n}.$$

Этот коэффициент принимает значение между 0 и 1. Большее значение R^2 соответствует лучшему подмножеству признаков.

Таблица 4: Обзор предлагаемых обобщений метода QPFS для векторной целевой переменной

Метод	Идея	Функция ошибки $S(\mathbf{z}\|\mathbf{X},\mathbf{Y})$
RelAgg	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$	$\min\limits_{\mathbf{z}_x}\big[(1-\alpha)\cdot\mathbf{z}_x^\intercal\mathbf{Q}_x\mathbf{z}_x - \alpha\cdot\mathbf{z}_x^\intercal\mathbf{B}\mathbf{1}_r\big]$
SymImp	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y}) + \mathrm{Sim}(\mathbf{Y})\big]$	$\min\limits_{\mathbf{z}_x,\,\mathbf{z}_y}\Big[\alpha_1\cdot\mathbf{z}_x^\intercal\mathbf{Q}_x\mathbf{z}_x - \alpha_2\cdot\mathbf{z}_x^\intercal\mathbf{B}\mathbf{z}_y + \alpha_3\cdot\mathbf{z}_y^\intercal\mathbf{Q}_y\mathbf{z}_y\Big]$
MinMax	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$ $\max\big[\mathrm{Rel}(\mathbf{X},\mathbf{Y}) + \mathrm{Sim}(\mathbf{Y})\big]$	$\min\limits_{\mathbf{z}_x}\max\limits_{\mathbf{z}_y}\big[\alpha_1\cdot\mathbf{z}_x^\intercal\mathbf{Q}_x\mathbf{z}_x - \alpha_2\cdot\mathbf{z}_x^\intercal\mathbf{B}\mathbf{z}_y - \alpha_3\cdot\mathbf{z}_y^\intercal\mathbf{Q}_y\mathbf{z}_y\big]$
AsymImp	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$ $\max\big[\mathrm{Rel}(\mathbf{X},\mathbf{Y}) + \mathrm{Sim}(\mathbf{Y})\big]$	$\min\limits_{\mathbf{z}_x,\mathbf{z}_y}\big[\alpha_1\cdot\mathbf{z}_x^\intercal\mathbf{Q}_x\mathbf{z}_x - \alpha_2\cdot\Big(\mathbf{z}_x^\intercal\mathbf{B}\mathbf{z}_y - \mathbf{b}^\intercal\mathbf{z}_y\Big) + \alpha_3\cdot\mathbf{z}_y^\intercal\mathbf{Q}_y\mathbf{z}_y\big]$

Нормированная среднеквадратичная ошибка (sRMSE) отображает качество прогнозирования модели. Оценка sRMSE считается на тренировочной и тестовой выборке.

$$\mathrm{sRMSE}(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a}) = \sqrt{\frac{\mathrm{MSE}(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a})}{\mathrm{MSE}(\mathbf{Y},\overline{\mathbf{Y}})}} = \frac{\|\mathbf{Y}-\widehat{\mathbf{Y}}_\mathbf{a}\|_2}{\|\mathbf{Y}-\overline{\mathbf{Y}}\|_2}.$$

Здесь $\widehat{\mathbf{Y}}_\mathbf{a} = \mathbf{X}_\mathbf{a}\boldsymbol{\Theta}_\mathbf{a}^\intercal$ — прогноз модели, $\overline{\mathbf{Y}}$ — предсказание константной модели, полученное усреднением целевой переменной по всем объектам. Данный показатель на тестовой выборке необходимо минимизировать.

Байесовский информационный критерий (BIC) — компромисс между качеством предсказания и размером выбранного подмножества признаков $\|\mathbf{a}\|_0 = \#\{j : a_j \neq 0\} = \sum_{j=1}^n a_j$:

$$\mathrm{BIC} = m\ln\Big(\mathrm{MSE}(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a})\Big) + \|\mathbf{a}\|_0\cdot\ln m.$$

Чем меньше значение BIC, тем лучше набор признаков.

Данные. Вычислительный эксперимент проводился на данных электрокортикограмм. Описание данных приведено в разделе .

На Рис. 3 показаны матрицы корреляций для исходных матриц $\mathbf{X}$ и $\mathbf{Y}$ данных ECoG. Частоты в матрице $\mathbf{X}$ сильно коррелированы. В целевой матрице $\mathbf{Y}$ корреляции между осями несущественны по сравнению с корреляциями между последовательными моментами времени и эти корреляции спадают со временем.

Результаты. Применим метод SymImp QPFS для различных значений коэффициента α_3 согласно формуле (30). Зависимость значимостей целевых векторов $\mathbf{z}_y$ относительно коэффициента α_3 для различных значений k показана на Рис. **??**. Значимости целевых векторов почти одинаковы для всех координат запястья при прогнозировании одного отсчёта времени ($k = 1$), что отражает

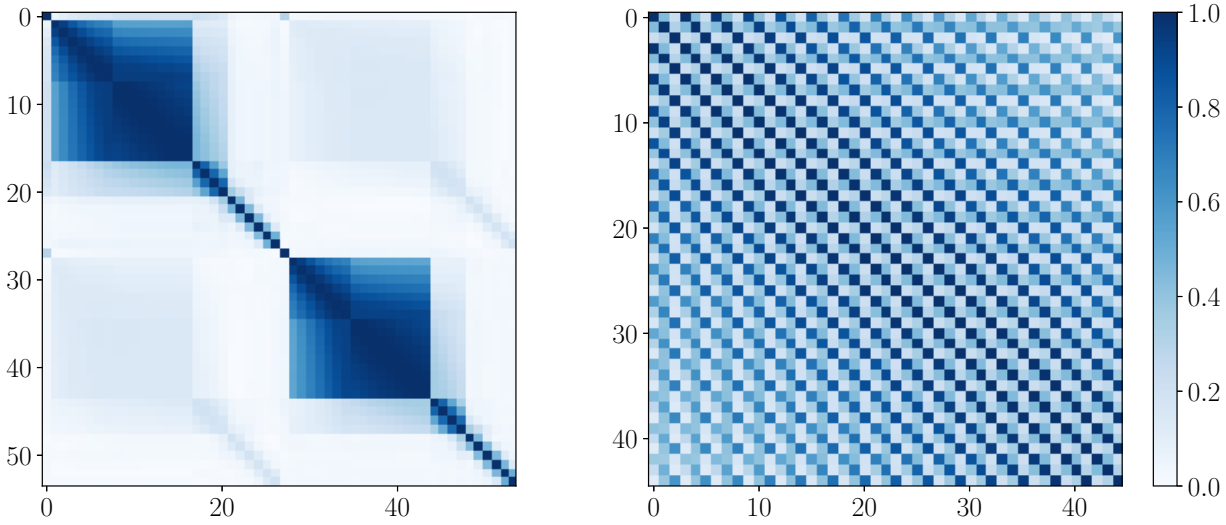

Рис. 3: Матрицы корреляций для матрицы плана $\mathbf{X}$ и целевой матрицы $\mathbf{Y}$ для данных ECoG

независимость между координатами x, y и z. Для $k = 2$ и $k = 3$ значимости некоторых целевых векторов становятся нулевыми при увеличении α_3. Вертикальные линии соответствуют оптимальному значению α_3, вычисленному по (29). При этом значении α_3 значимости компонент $\mathbf{z}_y$ совпадают. Таким образом, метод не учитывает различия между целевыми векторами для $k = 1, 2, 3$.

Предлагаемые методы QPFS для случая векторной целевой переменной, приведенные в таблице 4 применяются для набора данных ECoG. Решим задачу выбора признаков для каждого из методов, чтобы получить вектора значимостей признаков. Отсортируем по убыванию признаки по значению их значимостей. Обучим линейную модель, постепенно добавляя в неё признаки. Исследуются значения описанных критериев качества при увеличении количества отобранных признаков. На Рис. 4 показаны результаты для случая прогнозирования $k = 30$ отсчётов времени. Порог значимости признаков τ обозначен цветными тиками. Пороговые значения τ для предлагаемых методов больше, чем для базового метода RelAgg. Метод SymImp имеет большой порог, не позволяя получить малый набор признаков. Однако метод SymImp обладает наилучшей предсказательной способностью с точки зрения sRMSE на тестовых данных. Второй по качеству результат по sRMSE показал метод AsymImp. Все предложенные методы достигают меньшей ошибки на тестовой выборке по сравнению с методом RelAgg. Критерий устойчивости также выше для предложенных методов. Метод AsymImp показывает лучшие результаты с точки зрения качества прогнозирования и размера выбранного подмножества признаков.

Чтобы сравнить структуру выбранных подмножеств признаков и исследовать стабильность процедуры выбора признаков, используется метод генерации данных с помощью бутстрепа. Генерируется множество подвыборок, выбирая объекты по одному с возвращениями. Затем решается задача выбора признаков

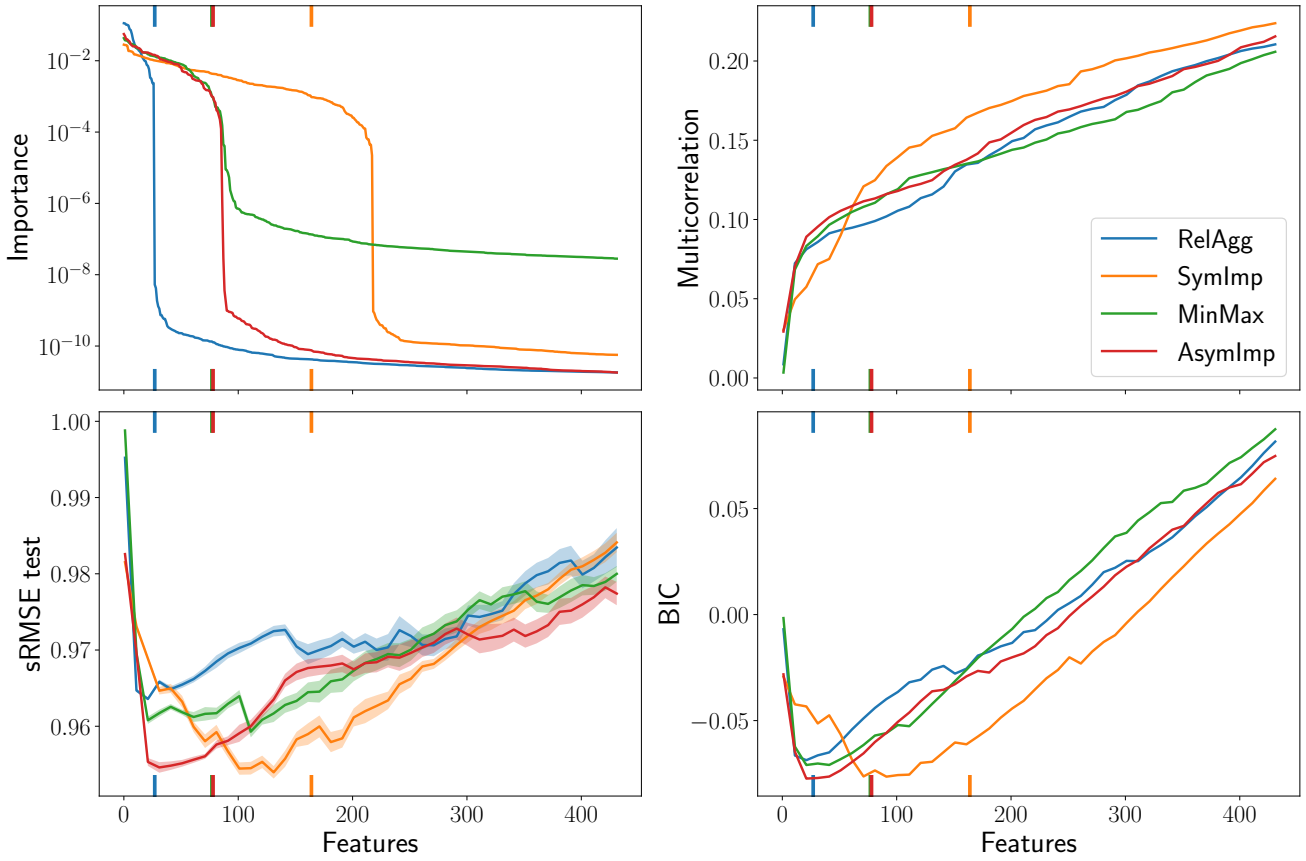

Рис. 4: Сравнение предложенных методов выбора признаков для данных ECoG при прогнозировании $k = 30$ отсчётов времени

для каждой пары матрицы исходных объектов $\mathbf{X}$ и матрицы целевых объектов $\mathbf{Y}$. Сравниваются полученные вектора значимостей для различных подвыборок данных. В качестве меры стабильности работы методов вычисляется средний попарный коэффициент корреляции Спирмена и попарное ℓ_2 расстояние. В таблице 5 показана средняя ошибка sRMSE, размер подмножества признаков и описанные статистики для каждого метода. Ошибка считалась на обученной линейной модели с использованием 50 признаков с наибольшими значениями значимостей. Asymimp дает наименьшую ошибку на тестовой выборке. Размер выбранных подмножеств объектов завышен при использовании порогового значения $\tau = 10^{-4}$. Оптимальное значение τ может быть подобрано с помощью процедуры кросс-валидации.

Таблица 5: Стабильность предложенных методов выбора признаков

	sRMSE	$\|\mathbf{a}\|_0$	Spearman ρ	ℓ_2
RelAgg	0.965 ± 0.002	26.8 ± 3.8	0.915 ± 0.016	0.145 ± 0.018
SymImp	0.961 ± 0.001	224.4 ± 9.0	0.910 ± 0.017	0.025 ± 0.002
MinMax	0.961 ± 0.002	101.0 ± 2.1	0.932 ± 0.009	0.059 ± 0.004
AsymImp	0.955 ± 0.001	85.8 ± 10.2	0.926 ± 0.011	0.078 ± 0.007

Для того, чтобы сравнить методы снижения размерности и выбора признаков, используется модель PLS, описанная в главе **??**. На Рис. 5 показана ошибка sRMSE на тренировочной и тестовой выборках в зависимости от размерности скрытого пространства l. Ошибка на тестовой выборке достигает минимума при $l = 11$. Метод PLS является более гибким подходом по сравнению с линейной моделью, построенной на подмножестве признаков, так как использует все исходные признаки. Это приводит к меньшей ошибке, но модель не является разреженной.

На Рис. 6 приведено сравнение 3 моделей: линейной регрессии; регрессии PLS, построенной на 100 признаках QPFS; регрессии PLS со всеми признаками. Линейная регрессия со всеми признаками не рассматривается, так как ее результаты близки к константному прогнозу. На рисунке также приведены результаты методов lasso и elastic net, которые широко используются для выбора признаков. В данном эксперименте использовался метод Asymimp QPFS. Размерность скрытого пространства PLS $l = 15$. Результаты регрессии PLS значительно лучше, линейной регрессии с признаками QPFS. Это означает, что последняя модель не является достаточно гибкой. Тем не менее, лучший результат показывает модель PLS, построенная на признаках QPFS. Данная модель является разреженной, так как использует только 100 исходных признаков. Способность модели PLS находить оптимальное скрытое представление данных улучшает предсказательную способность модели.

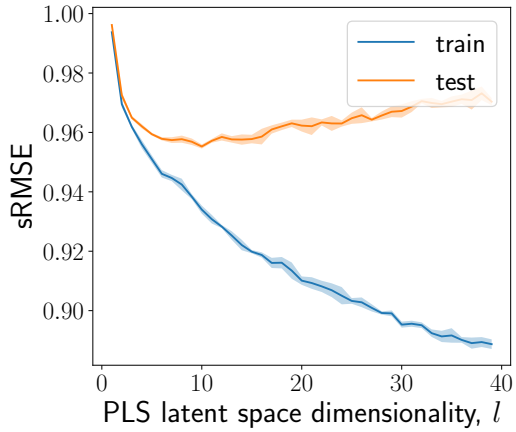

Рис. 5: Ошибка sRMSE на тестовой выборке для модели PLS

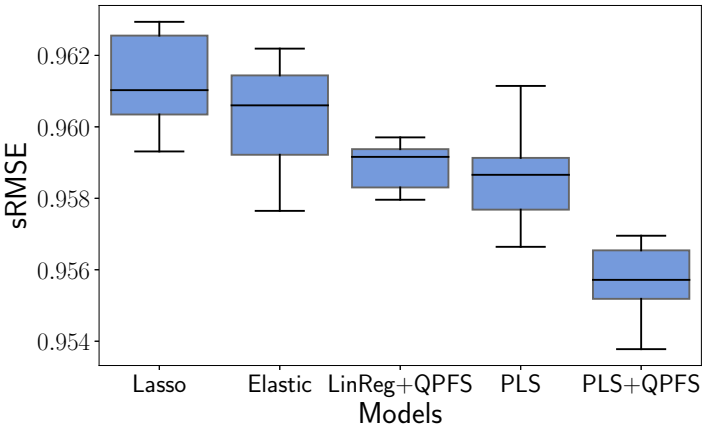

Рис. 6: Диаграммы размаха значений sRMSE на тестовой выборке для моделей Lasso, Elastic, LinReg+QPFS, PLS, PLS+QPFS

Глава 4. Выбор параметров нелинейных моделей с помощью квадратичного отбора признаков

Функция ошибки для моделей с большим числом параметров имеет сложный ландшафт с многими локальными минимумами. В этом случае алгоритм оптимизации приводит к разным решениям в зависимости от инициализации исходных параметров.

Алгоритм оптимизации представляет собой итерационный процесс. На каждом шаге для получения следующего приближения параметров модели обновляются текущие параметры. Разработано множество алгоритмов оптимизации первого порядка, использующих вектор первых производных функции ошибки. Наиболее известными алгоритмами являются градиентный спуск, метод момента Нестерова [?], AdaGrad [?], Adam [?]. Данные алгоритмы используются для оптимизации глубоких нейронных сетей [?]. Метод Ньютона — алгоритм второго порядка, использующий матрицу вторых производных функции ошибки. Метод Ньютона находит обновления параметров для квадратичной аппроксимации функции ошибки и сходится за адекватное число итераций. Недостатком методов оптимизации второго порядка является огромная и плохо обусловленная матрица Гессиана. Процесс оптимизации в этом случае расходится и является вычислительно дорогостоящим. Авторы [?, ?] предлагают аппроксимации для матрицы Гессиана и регуляризацию для решения этой проблемы. В статье [?] метод Ньютона применяется к глубоким нейронным сетям.

В данной главе приводится анализ параметров модели, которые не находятся в оптимуме. Приводится метод выбора активных параметров модели, основанный на методе QPFS, который подробно описан в главе 0.1. Рассматриваются задачи нелинейной регрессии с квадратичной функцией потерь, логистической регрессии с кросс-энтропийной функцией потерь.

Задача выбора параметров для оптимизации нелинейных моделей.

Модель $f(\mathbf{x}, \boldsymbol{\theta})$ с параметрами $\boldsymbol{\theta} \in \mathbb{R}^p$ предсказывает целевой объект $y \in \mathbb{Y}$ по исходному объекту $\mathbf{x} \in \mathbb{R}^n$. Пространство $\mathbb{Y}$ представляет собой бинарные метки классов $\{0, 1\}$ для задачи двухклассовой классификации и $\mathbb{R}$ для задачи регрессии. Даны матрица исходных объектов $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\intercal \in \mathbb{R}^{m \times n}$ и целевой вектор $\mathbf{y} = [y_1, \ldots, y_m]^\intercal \in \mathbb{Y}^m$. Цель состоит в нахождении оптимальных параметров $\boldsymbol{\theta}^*$. Параметры $\boldsymbol{\theta}$ вычисляются минимизацией функции ошибки:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}). \tag{42}$$

Данная задача полностью соответствует рассмотренной задаче (3) для случая скалярной целевой переменной ($r = 1$). В качестве функции ошибки $\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$ рассматриваются квадратичная ошибка для задачи регрессии:

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) = \frac{1}{2}\|\mathbf{y} - \mathbf{f}(\mathbf{X}, \boldsymbol{\theta})\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}\big(y_i - f(\mathbf{x}_i, \boldsymbol{\theta})\big)^2, \tag{43}$$

и функция кросс-энтропии для задачи бинарной классификации:

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{m}\big[y_i \log f(\mathbf{x}_i, \boldsymbol{\theta}) + (1 - y_i)\log\big(1 - f(\mathbf{x}_i, \boldsymbol{\theta})\big)\big]. \tag{44}$$

Задача (42) решается с помощью итеративной процедуры оптимизации. Для получения параметров на шаге k текущие параметры $\boldsymbol{\theta}^{k-1}$ обновляются по следующему правилу:

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} + \Delta\boldsymbol{\theta}^{k-1}. \tag{45}$$

Авторы используют метод оптимизации Ньютона для выбора вектора обновлений $\Delta\boldsymbol{\theta}$.

Метод Ньютона нестабилен и вычислительно сложен. В данной статье предлагается стабильный метод Ньютона. Перед шагом градиента предлагается выбрать подмножество активных параметров модели, которые оказывают наибольшее влияние на функцию ошибки $\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$. Введём определение активного параметра модели, используя необходимое условие оптимальности первого порядка.

Определение 10. Параметр θ_j для модели $f(\mathbf{x}, \boldsymbol{\theta})$ является *активным*, если $\mathbf{J}^\intercal(\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{y}) \neq 0$.

Подробный вывод условия из определения приводится в разделе 0.2. Обновление параметров производится только для отобранного множества индексов $\mathcal{A} = \left\{j : a_j = 1, \mathbf{a} \in \{0,1\}^p\right\}$

$$\begin{aligned}
&\boldsymbol{\theta}^k_{\mathcal{A}} = \boldsymbol{\theta}^{k-1}_{\mathcal{A}} + \Delta\boldsymbol{\theta}^{k-1}_{\mathcal{A}}, \quad \boldsymbol{\theta}_{\mathcal{A}} = \{\theta_j : j \in \mathcal{A}\},\\
&\boldsymbol{\theta}^k_{\bar{\mathcal{A}}} = \boldsymbol{\theta}^{k-1}_{\bar{\mathcal{A}}}, \quad \boldsymbol{\theta}_{\bar{\mathcal{A}}} = \{\theta_j : j \notin \mathcal{A}\}.
\end{aligned}$$

Чтобы выбрать оптимальное подмножество индексов $\mathcal{A}$, из всех возможных $2^p - 1$ подмножеств, вводится функция ошибки

$$\mathbf{a} = \underset{\mathbf{a}' \in \{0,1\}^p}{\arg\min}\, S(\mathbf{a}', \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}), \tag{46}$$

аналогичная функции ошибки (21) для задачи выбора признаков. Задача (46) решается на каждом шаге k процесса оптимизации для текущих параметров $\boldsymbol{\theta}^k$.

Метод QPFS используется для решения задачи (46). QPFS выбирает подмножество параметров $\mathbf{a}$ для вектора обновлений $\Delta\boldsymbol{\theta}$, которые оказывают наибольшее влияние на вектор остатков и являются попарно независимыми. Функция ошибки (25) соответствует функции ошибки $S(\mathbf{a}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta})$

$$\mathbf{a} = \underset{\mathbf{a}' \in \{1,0\}^p}{\arg\max}\, S(\mathbf{a}', \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) \Leftrightarrow \underset{\mathbf{a} \in \mathbb{R}^p_+,\, \|\mathbf{a}\|_1 = 1}{\arg\min} \left[\mathbf{a}^\intercal \mathbf{Q} \mathbf{a} - \alpha \cdot \mathbf{b}^\intercal \mathbf{a}\right]. \tag{47}$$

В работе показано, что для модели нелинейной регрессии с квадратичной функцией ошибки (43) и для модели логистической регрессии с кросс-энтропией (44), каждый шаг оптимизации эквивалентен задаче линейной регрессии (24).

Метод Ньютона для оптимизации параметров.

Метод Ньютона использует условие оптимизации первого порядка для задачи (42) и линеаризует градиент $S(\boldsymbol{\theta})$

$$\nabla S(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \nabla S(\boldsymbol{\theta}) + \mathbf{H} \cdot \Delta\boldsymbol{\theta} = 0,$$

$$\Delta\boldsymbol{\theta} = -\mathbf{H}^{-1}\nabla S(\boldsymbol{\theta}).$$

где $\mathbf{H} = \nabla^2 S(\boldsymbol{\theta})$ является матрицей Гессиана функции ошибки $S(\boldsymbol{\theta})$.

Итерация (45) метода Ньютона имеет вид

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} - \mathbf{H}^{-1}\nabla S(\boldsymbol{\theta}).$$

На каждой итерации требуется обращать матрицу Гессиана $\mathbf{H}$. Мерой плохой обусловленности для матрицы Гессиана $\mathbf{H}$ является число обусловленности

$$\varkappa(\mathbf{H}) = \frac{\lambda_{\max}(\mathbf{H})}{\lambda_{\min}(\mathbf{H})},$$

где $\lambda_{\max}(\mathbf{H}), \lambda_{\min}(\mathbf{H})$ являются максимальным и минимальным собственными значениями $\mathbf{H}$. Большое число обусловленности $\varkappa(\mathbf{H})$ приводит к нестабильности процесса оптимизации. Предложенный метод уменьшает размер матрицы Гессиана $\mathbf{H}$. Согласно экспериментам, приведенным в разделе 0.2 предлагаемый метод приводит к меньшему числу обусловленности $\varkappa(\mathbf{H})$.

Размер шага метода Ньютона может быть чрезмерно большим. Для контроля размера шага обновлений добавим параметр η в правило обновления (45)

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} + \eta\Delta\boldsymbol{\theta}^{k-1}, \quad \eta \in [0, 1].$$

Для выбора соответствующего размера шага η используется правило Армихо [**?**]. Выбирается максимальное η так, чтобы выполнялось условие

$$S(\boldsymbol{\theta}^{k-1} + \eta\Delta\boldsymbol{\theta}^{k-1}) < S(\boldsymbol{\theta}^{k-1}) + \gamma\eta\nabla S^{\mathsf{T}}(\boldsymbol{\theta}^{k-1})\boldsymbol{\theta}^{k-1}, \quad \gamma \in [0, 0.5].$$

Модель нелинейной регрессии. Предположим, что модель $f(\mathbf{x}, \boldsymbol{\theta})$ близка к линейной в окрестности точки $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$

$$\mathbf{f}(\mathbf{X}, \boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \approx \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) + \mathbf{J} \cdot \Delta\boldsymbol{\theta},$$

где $\mathbf{J} \in \mathbb{R}^{m\times p}$ является матрицей Якоби

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f(\mathbf{x}_1,\boldsymbol{\theta})}{\partial\theta_1} & \cdots & \frac{\partial f(\mathbf{x}_1,\boldsymbol{\theta})}{\partial\theta_p} \\ \cdots & \cdots & \cdots \\ \frac{\partial f(\mathbf{x}_m,\boldsymbol{\theta})}{\partial\theta_1} & \cdots & \frac{\partial f(\mathbf{x}_m,\boldsymbol{\theta})}{\partial\theta_p} \end{pmatrix}.$$

В соответствии с этим предположением градиент $\nabla S(\boldsymbol{\theta})$ и Гессиан матрицы $\mathbf{H}$ функции ошибки (43) равняются

$$\nabla S(\boldsymbol{\theta}) = \mathbf{J}^{\mathsf{T}}(\mathbf{y} - \mathbf{f}), \quad \mathbf{H} = \mathbf{J}^{\mathsf{T}}\mathbf{J}. \tag{48}$$

Данные предположения приводят к методу Гаусса-Ньютона и правилу обновления (45)

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} + (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}\mathbf{J}^{\mathsf{T}}(\mathbf{f} - \mathbf{y}).$$

Вектор обновления $\Delta\boldsymbol{\theta}$ является решением задачи линейной регрессии

$$\|\mathbf{e} - \mathbf{F}\Delta\boldsymbol{\theta}\|_2^2 \to \min_{\Delta\boldsymbol{\theta}\in\mathbb{R}^p}, \tag{49}$$

где $\mathbf{e} = \mathbf{f} - \mathbf{y}$ и $\mathbf{F} = \mathbf{J}$.

В качестве нелинейной модели рассматривается модель двухслойной нейронной сети. В этом случае модель $f(\mathbf{x}, \boldsymbol{\theta})$ принимает вид:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sigma(\mathbf{x}^\intercal \mathbf{W}_1)\mathbf{w}_2.$$

Здесь $\mathbf{W}_1 \in \mathbb{R}^{m\times h}$ — матрица параметров, которые соединяют исходные признаки с h скрытыми нейронами. Функция нелинейности $\sigma(\cdot)$ применяется поэлементно. Параметры $\mathbf{w}_2 \in \mathbb{R}^{h\times 1}$ соединяют скрытые нейроны с выходом. Вектор параметров модели $\boldsymbol{\theta}$ представляет собой объединение векторизованных матриц $\mathbf{W}_1$, $\mathbf{w}_2$.

Модель логистической регрессии. Для логистической регрессии модель имеет вид $f(\mathbf{x}, \boldsymbol{\theta}) = \sigma(\mathbf{x}^\intercal \boldsymbol{\theta})$ с сигмоидной функцией активации $\sigma(\cdot)$. Градиент и Гессиан функции ошибки (44) равны

$$\nabla S(\boldsymbol{\theta}) = \mathbf{X}^\intercal(\mathbf{f} - \mathbf{y}), \quad \mathbf{H} = \mathbf{X}^\intercal\mathbf{R}\mathbf{X}, \tag{50}$$

где $\mathbf{R}$ — это диагональная матрица с диагональными элементами $f(\mathbf{x}_i, \boldsymbol{\theta}) \cdot (1 - f(\mathbf{x}_i, \boldsymbol{\theta}))$.

Правило обновления (45) в этом случае принимает вид

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} + (\mathbf{X}^\intercal\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{y} - \mathbf{f}).$$

Этот алгоритм известен как итеративный алгоритм взвешенных наименьших квадратов (IRLS) [**?**]. Вектор обновлений $\Delta\boldsymbol{\theta}$ является решением задачи линейной регрессии

$$\|\mathbf{e} - \mathbf{F}\Delta\boldsymbol{\theta}\|_2^2 \to \min_{\Delta\boldsymbol{\theta}\in\mathbb{R}^p}, \tag{51}$$

где $\mathbf{e} = \mathbf{R}^{-1/2}(\mathbf{y} - \mathbf{f})$ и $\mathbf{F} = \mathbf{R}^{1/2}\mathbf{X}$.

0.2 Метод Ньютона с выбором параметров с помощью квадратичного программирования

Предлагается адаптировать метод QPFS для решения задач (49) и (51). Матрица парных взаимодействий $\mathbf{Q}$ и вектор релевантностей $\mathbf{b}$ имеют вид

$$\mathbf{Q} = \mathrm{Sim}(\mathbf{F}), \quad \mathbf{b} = \mathrm{Rel}(\mathbf{F}, \mathbf{e}).$$

Выборочный коэффициент корреляции равен нулю для ортогональных векторов. Покажем, что в оптимальной точке $\boldsymbol{\theta}^*$ вектор $\mathbf{e}$ ортогонален столбцам матрицы $\mathbf{F}$. В этом случае вектор $\mathbf{b} = \mathrm{Rel}(\mathbf{F}, \mathbf{e})$ равен нулю. Это означает, что

член, учитывающий релевантность, в данном случае исключается. Условие оптимальности первого порядка гарантирует это свойство для модели нелинейной регрессии

$$\mathbf{F}^\intercal \mathbf{e} = \mathbf{J}^\intercal(\mathbf{f} - \mathbf{y}) = -\nabla S(\boldsymbol{\theta}^*) = \mathbf{0},$$

и для модели логистической регрессии

$$\mathbf{F}^\intercal \mathbf{e} = \mathbf{X}\mathbf{R}^{-1/2}\mathbf{R}^{1/2}(\mathbf{y} - \mathbf{f}) = \mathbf{X}^\intercal(\mathbf{y} - \mathbf{f}) = \nabla S(\boldsymbol{\theta}^*) = \mathbf{0}.$$

Данное условие используется в качестве индикатора активности параметра модели в определении 10. Псевдокод предлагаемого алгоритма приведён в алгоритме 2.

Algorithm 2 QPFS + Ньютон алгоритм

Вход: ε — допустимое отклонение;
τ – пороговое значение;
γ — параметр правила Армихо.
Выход: $\boldsymbol{\theta}^*$;
инициализировать $\boldsymbol{\theta}^0$;
$k := 1$;
повторять
вычислить $\mathbf{e}$ и $\mathbf{F}$ для (49) или (51) ;
$\mathbf{Q} := \mathrm{Sim}(\mathbf{F})$, $\mathbf{b} := \mathrm{Rel}(\mathbf{F}, \mathbf{e})$, $\alpha = \frac{\overline{\mathbf{Q}}}{\overline{\mathbf{Q}+\mathbf{b}}}$;
$\mathbf{a} := \underset{\mathbf{a}\geqslant 0,\, \|\mathbf{a}\|_1=1}{\arg\min}\ \mathbf{a}^\intercal\mathbf{Q}\mathbf{a} - \alpha \cdot \mathbf{b}^\intercal\mathbf{a}$;
$\mathcal{A} := \{j : a_j = 1\}$;
вычислить $\nabla S(\boldsymbol{\theta}^{k-1})$, $\mathbf{H}$ для (48) или (50);
$\Delta\boldsymbol{\theta}^{k-1} = -\mathbf{H}^{-1}\nabla S(\boldsymbol{\theta}^{k-1})$;
$\eta := \mathrm{ArmijoRule}(\boldsymbol{\theta}^{k-1}, \gamma)$;
$\boldsymbol{\theta}^k_{\mathcal{A}} = \boldsymbol{\theta}^{k-1}_{\mathcal{A}} + \eta\Delta\boldsymbol{\theta}^{k-1}_{\mathcal{A}}$;
$k := k + 1$;
пока $\frac{\|\boldsymbol{\theta}^k - \boldsymbol{\theta}^{k-1}\|}{\|\boldsymbol{\theta}^k\|} < \varepsilon$

Анализ значимостей параметров нелинейных моделей.

Целью вычислительного эксперимента является исследование свойств предложенного метода и сравнение его с другими методами.

Исследована зависимость параметров метода QPFS для задачи нелинейной регрессии (49) и задачи логистической регрессии (51). Предположим, что вектор параметров $\boldsymbol{\theta}^0$ лежит вблизи оптимального вектора параметров $\boldsymbol{\theta}^*$. Рассмотрим отрезок

$$\boldsymbol{\theta}_\beta = \beta\boldsymbol{\theta}^* + (1 - \beta)\boldsymbol{\theta}^0; \quad \beta \in [0, 1].$$

Сгенерируем синтетический набор данных с 300 объектами и 7 признаками для задачи логистической регрессии. Ландшафт функции ошибки (44) на сетке двух случайно выбранных параметров показан на Рис. **??**. Поверхность функции ошибки выпуклая с вытянутыми линиями уровня вдоль некоторых параметров модели. Добавим случайный шум к оптимальным параметрам $\boldsymbol{\theta}^*$, чтобы получить точку $\boldsymbol{\theta}^0$. Поведение вектора $\mathbf{b}$ на отрезке между $\boldsymbol{\theta}^0$ и $\boldsymbol{\theta}^*$ показано на Рис. **??**. Компоненты $\mathbf{b}$ начинают резко уменьшаться по мере приближения к оптимальной точке $\boldsymbol{\theta}^*$.

Для модели нелинейной регрессии используется классический набор данных Boston Housing с 506 объектами и 13 признаками. Для простоты нейронная сеть содержит два скрытых нейрона. Ландшафт функции ошибок для модели нейронной сети является более сложным. Функция ошибки не является выпуклой и содержит множество локальных минимумов. Двумерный ландшафт функции ошибок для этого набора данных показан на Рис. **??**. Сетка строится для двух случайных параметров из матрицы $\mathbf{W}_1$. Аналогично на Рис. **??** показано, как изменяется вектор $\mathbf{b}$ при движении от точки $\boldsymbol{\theta}^0$ до точки $\boldsymbol{\theta}^*$. Компоненты вектора $\mathbf{b}$ становятся близки к нулю вблизи оптимума. При достижении оптимального значения различные параметры влияют на остатки модели $\mathbf{e}$.

На Рис. **??** показан процесс оптимизации для предложенного метода в случае логистической регрессии с двумя параметрами модели. Даже для двумерной задачи решение метода Ньютона нестабильно и число обусловленности $\varkappa(\mathbf{H})$ матрицы Гессиана $\mathbf{H}$ может быть чрезвычайно большим. На каждом шаге алгоритма метод QPFS выбирает активные параметры для оптимизации. В данном примере предложенный метод выбирает и обновляет только один параметр на каждой итерации на первых шагах. Это делает метод более устойчивым.

На Рис. **??** показаны наборы активных параметров на итерациях для набора данных Boston Housing и нейронной сети с двумя скрытыми нейронами. Темные ячейки соответствуют активным параметрам, которые мы оптимизируем.

В рассмотренных примерах число обусловленности $\varkappa(\mathbf{H})$ для метода Ньютона на некоторых итерациях было чрезвычайно большим. Выбор активных параметров позволил значительно сократить число обусловленности.

Приведём сравнение предложенного метода с существующими методами, а именно градиентным спуском (GD), моментом Нестерова [**?**], Adam [**?**] и оригинальным методом Ньютона. Проведены эксперименты для моделей нелинейной и логистической регрессий. Наборы данных были выбраны из репозитория UCI [**?**]. Результаты показаны в таблицах 6 и 7. Для каждого набора данных две строки таблиц содержат ошибки для тренировочной (первая строка) и тестовой (вторая строка) выборок. В таблице 6 приведена квадратичная ошибка, в таблице 7 — кросс-энтропия. Чтобы найти среднюю ошибку и ее стандартное отклонение использовалась процедура кросс валидации с разбиением на 5 фолдов. Предложенный метод показывает меньшую ошибку на трех из четырех наборов данных для нелинейной регрессии и на двух из трех наборов данных для логистической регрессии.

Таблица 6: Средняя квадратичная ошибка рассматриваемых алгоритмов оптимизации для модели нелинейной регрессии

Выборка	m / n	GD	Нестеров	ADAM	Ньютон	QPFS+Ньютон
Boston House Prices	506 / 13	27.2 ± 4.6 / 32.4 ± 5.6	46.0 ± 11.0 / 53.3 ± 11.5	35.4 ± 2.5 / 37.8 ± 7.0	22.1 ± 15.2 / 28.9 ± 13.6	20.9 ± 10.4 / $\mathbf{24.5 \pm 9.4}$
Communities and Crime	1994 / 99	48.0 ± 6.4 / $47{,}5 \pm 6.5$	31.4 ± 2.8 / 32.9 ± 4.3	23.3 ± 3.7 / $28{,}1 \pm 4.5$	18.3 ± 3.4 / 28.8 ± 3.6	26.7 ± 3.1 / $\mathbf{28.4 \pm 3.0}$
Forest Fires	517 / 10	18.9 ± 0.4 / $\mathbf{20.0 \pm 2.1}$	1.83 ± 0.4 / 20.2 ± 2.2	1.81 ± 0.6 / $\mathbf{20.0 \pm 2.0}$	17.7 ± 0.4 / 20.6 ± 1.4	17.9 ± 0.4 / 20.2 ± 2.2
Residential Building	372 / 103	51.6 ± 17.7 / 53.7 ± 13.9	32.6 ± 19.5 / 34.1 ± 13.6	30.0 ± 24.8 / 34.1 ± 19.4	35.5 ± 24.7 / 35.0 ± 15.6	30.3 ± 10.7 / $\mathbf{30.9 \pm 5.3}$

Таблица 7: Среднее значение кросс-энтропии рассматриваемых алгоритмов оптимизации для модели логистической регрессии

Выборка	m / n	GD	Нестеров	ADAM	Ньютон	QPFS+Ньютон
Breast Cancer	569 / 30	0.6 ± 0.1 / $\mathbf{0.9 \pm 0.2}$	0.4 ± 0.1 / 1.0 ± 0.7	0.8 ± 0.2 / 1.2 ± 0.2	0.3 ± 0.1 / 1.0 ± 0.2	0.2 ± 0.1 / 1.1 ± 0.3
Cardiotocography	2126 / 21	11.5 ± 4.7 / 11.6 ± 5.8	11.5 ± 4.7 / 11.5 ± 5.7	8.8 ± 4.4 / 9.0 ± 2.6	11.5 ± 5.7 / 11.5 ± 4.7	7.7 ± 4.2 / $\mathbf{7.7 \pm 4.7}$
Climate Model Simulation Crashes	540 / 18	1.2 ± 0.1 / 1.4 ± 2.0	1.0 ± 0.2 / 1.3 ± 0.7	1.5 ± 0.2 / 1.8 ± 0.3	1.0 ± 0.5 / 1.2 ± 0.5	0.8 ± 0.3 / $\mathbf{1.1 \pm 0.4}$

Глава 5. Метрические методы анализа временных рядов

При использовании в качестве функции ошибки модели квадратичной ошибки предполагается, что целевое пространство является евклидовым. Данное предположение не всегда является адекватным. В данной главе ставится задача метрического обучения как поиск оптимальной метрики в целевом пространстве. Рассматриваются задачи кластеризации и классификации множества временных рядов.

Метрическое обучение в задачах кластеризации временных рядов.

Пусть $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\intercal \in \mathbb{R}^{m \times n}$ — матрица плана. Объект $\mathbf{x}_i = [x_i^1, \ldots, x_i^n]^\intercal$ задан в виде вектора в пространстве признаков. Требуется выявить кластерную структуру данных и разбить множество объектов $\mathbf{X}$ на множество непересекающихся кластеров $\mathbb{Y} = \{1, \ldots, K\}$, т. е. построить отображение $f : \mathbb{R}^n \to \mathbb{Y}$. Обозначим $y_i = f(\mathbf{x}_i)$, $y_i \in \mathbb{Y}$ — метка кластера объекта $\mathbf{x}_i$. Необходимо выбрать метки кластеров $\{y_i\}_{i=1}^m$ таким образом, чтобы расстояния между кластерами были максимальными. Центроид $\boldsymbol{\mu}$ множества объектов $\mathbf{X}$ и центроиды кластеров $\{\boldsymbol{\mu}_j\}_{j=1}^K$ вычисляются по формулам:

$$\boldsymbol{\mu} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i, \quad \boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m} [y_i = y_j] \mathbf{x}_i}{\sum_{i=1}^{m} [y_i = y_j]}. \tag{52}$$

Введем на множестве объектов $\mathbf{X}$ расстояние Махаланобиса

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A}^{-1}(\mathbf{x}_i - \mathbf{x}_j)}, \tag{53}$$

где *матрица трансформаций* $\mathbf{A} \in \mathbb{R}^{n \times n}$ является симметричной и неотрицательно определенной ($\mathbf{A}^\intercal = \mathbf{A}$, $\mathbf{A} \succeq 0$). Зададим в качестве матрицы трансформации матрицу выборочной ковариации

$$\mathbf{A} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\intercal. \tag{54}$$

Функцией ошибки кластеризации назовем межкластерное расстояние:

$$\mathcal{L}\big(\{\boldsymbol{\mu}_j\}_{j=1}^K, \mathbf{X}, \mathbf{y}\big) = -\sum_{j=1}^{K} N_j d^2_{\mathbf{A}}(\boldsymbol{\mu}_j, \boldsymbol{\mu}), \tag{55}$$

где $N_j = \sum_{i=1}^m [y_i = y_j]$ — число объектов в кластере j.

Поставим задачу кластеризации как задачу минимизации функции ошибки (55)

$$\mathcal{L}\big(\{\boldsymbol{\mu}_j\}_{j=1}^K, \mathbf{X}, \mathbf{y}\big) \to \min_{\boldsymbol{\mu}_j \in \mathbb{R}^\intercal}. \tag{56}$$

Для решения этой задачи предлагается применить метод метрического обучения к матрице трансформации $\mathbf{A}$. Найдем такую матрицу $\mathbf{A}$, для которой функционал качества принимает максимальное значение:

$$\mathbf{A}^* = \arg\min_{\mathbf{A} \in \mathbb{R}^{n \times n}} S\big(\{\boldsymbol{\mu}_j^*\}_{j=1}^K, \mathbf{X}, \mathbf{y}\big), \tag{57}$$

где $\{\boldsymbol{\mu}_j^*\}_{j=1}^K$ — решение задачи кластеризации (56).

Алгоритм адаптивного метрического обучения.

Для решения задач (56), (57) используется алгоритм адаптивного метрического обучения. Предлагается понизить размерность пространства объектов $\mathbf{X}$ с помощью линейного ортогонального преобразования $\mathbf{P} \in \mathbb{R}^{l \times n}$, $\mathbf{P}^\intercal \mathbf{P} = \mathbf{I}$, где новая размерность $l < n$

$$\mathbf{t}_i = \mathbf{P}\mathbf{x}_i \in \mathbb{R}^l, \quad i = 1, \ldots, m.$$

Центроид $\hat{\boldsymbol{\mu}}$ множества объектов $\{\mathbf{t}_i\}_{i=1}^m$ вычисляется по формуле (52). Расстояния между объектами вычисляются по формуле (53), где в качестве матрицы $\hat{\mathbf{A}}$ используется матрица ковариаций (54) множества объектов $\{\hat{\mathbf{x}}_i\}_{i=1}^m$

$$\hat{\mathbf{A}} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{t}_i - \hat{\boldsymbol{\mu}})(\mathbf{t}_i - \hat{\boldsymbol{\mu}})^\intercal = \frac{1}{m} \sum_{i=1}^{m} \mathbf{P}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\intercal \mathbf{P}^\intercal = \mathbf{P}\mathbf{A}\mathbf{P}^\intercal.$$

Определение 11. Индикаторной матрицей назовем матрицу $\mathbf{Y} \in \mathbb{R}^{m \times K}$, где

$$y_{ij} = \begin{cases} 1, & \text{если } f(\mathbf{x}_i) = y_j; \\ 0, & \text{если } f(\mathbf{x}_i) \neq y_j. \end{cases}$$

Определение 12. Взвешенной индикаторной матрицей назовем матрицу $\mathbf{L} = \mathbf{Y}(\mathbf{Y}^\intercal\mathbf{Y})^{-1/2} \in \mathbb{R}^{m \times K}$, элементы которой равны:

$$l_{ij} = \begin{cases} \dfrac{1}{\sqrt{N_j}}, & \text{если } f(\mathbf{x}_i) = y_j; \\ 0, & \text{если } f(\mathbf{x}_i) \neq y_j. \end{cases}$$

В работе [?] показано, что с использованием данных обозначений задача кластеризации (56) и задача метрического обучения (57) сводятся к общей задаче минимизации функции ошибки

$$\mathcal{L} = -\frac{1}{m}\text{trace}(\mathbf{L}^\intercal\mathbf{X}^\intercal\mathbf{P}^\intercal\hat{\mathbf{A}}^{-1}\mathbf{P}\mathbf{X}\mathbf{L}) = \\ = -\frac{1}{m}\text{trace}(\mathbf{L}^\intercal\mathbf{X}^\intercal\mathbf{P}^\intercal(\mathbf{P}\mathbf{A}\mathbf{P}^\intercal)^{-1}\mathbf{P}\mathbf{X}\mathbf{L}) \to \min_{\mathbf{P},\mathbf{L}}. \quad (58)$$

Для решения задачи (58) используется EM алгоритм. На каждом шаге итеративно вычисляются текущие оптимальные значения матриц $\mathbf{P}$ и $\mathbf{L}$. На E-шаге необходимо найти матрицу $\mathbf{L}$, которая является решением оптимизационной задачи (58) при фиксированной матрице $\mathbf{P}$. В качестве начального приближения получим взвешенную индикаторную матрицу $\mathbf{L}$ с помощью алгоритма кластеризации k-средних с евклидовой метрикой. На M-шаге производится нахождение оптимального значения матрицы $\mathbf{P}$ при фиксированной матрице $\mathbf{L}$. Алгоритм завершается при стабилизации функционала $\mathcal{L}$ на последовательности итераций.

Алгоритм k-средних. В данной работе базовым алгоритмом для сравнения является алгоритм k-средних. На первом шаге алгоритм выбирает из множества $\mathbf{X}$ случайным образом r объектов $\{\boldsymbol{\mu}_j\}_{j=1}^K$ — начальные центроиды кластеров. Для каждого объекта $\mathbf{x}_i$ вычисляется расстояние (53) до каждого центроида кластера $\boldsymbol{\mu}_j$ с единичной матрицей трансформаций $\mathbf{A}$. Объект $\mathbf{x}_i$ относится к кластеру, расстояние до которого оказалось наименьшим. Далее производится вычисление новых центроидов кластеров по формуле (52). Алгоритм завершается, если значения центроидов кластеров стабилизируются.

Оптимизация матрицы P с фиксированной матрицей L. Для любых двух квадратных матриц $\mathbf{A}$ и $\mathbf{B}$ справедливо $\text{trace}(\mathbf{A}\mathbf{B}) = \text{trace}(\mathbf{B}\mathbf{A})$. Данное свойство позволяет переформулировать задачу (58) следующим образом:

$$\mathcal{L} = -\frac{1}{m}\text{trace}(\mathbf{L}^\intercal\mathbf{X}^\intercal\mathbf{P}^\intercal(\mathbf{P}\mathbf{A}\mathbf{P}^\intercal)^{-1}\mathbf{P}\mathbf{X}\mathbf{L}) = -\frac{1}{m}\text{trace}\big((\mathbf{P}\mathbf{A}\mathbf{P}^\intercal)^{-1}\mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}^T\mathbf{X}^\intercal\mathbf{P}^\intercal\big).$$

Утверждение 11. Обозначим $\mathbf{B} = \mathbf{X}\mathbf{L}\mathbf{L}^{\intercal}\mathbf{X}^{\intercal}$. Обозначим через $\mathbf{P} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]^{\mathbf{T}}$ матрицу, состоящую из r собственных векторов матрицы $\mathbf{A}^{-1}\mathbf{B}$, отвечающих наибольшим собственным значениям. Тогда решением (58) является ортогональная матрица, полученная QR-разложением матрицы $\mathbf{P}^{\intercal}$.

Доказательство. Функция ошибки $\mathcal{L}$ 55 зависит только от матрицы $\mathbf{P}$. Обозначим

$$s(\mathbf{P}) = \operatorname{trace}\big((\mathbf{P}\mathbf{A}\mathbf{P}^{\intercal})^{-1}\mathbf{P}\mathbf{B}\mathbf{P}^{\intercal}\big).$$

На данном шаге задача (58) принимает вид:

$$\mathbf{P}^* = \arg\max_{\mathbf{P}\in\mathbb{R}^{l\times n}} s(\mathbf{P})\,; \tag{59}$$

$$\mathbf{P}\mathbf{P}^{\intercal} = \mathbf{I}\,. \tag{60}$$

Ранг произведения матриц не превосходит рангов сомножителей, поэтому ранг матрицы $\mathbf{B}$ не превосходит K. Решением (59) является матрица $\mathbf{P} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]^{\intercal}$, состоящая из K собственных векторов матрицы $\mathbf{A}^{-1}\mathbf{B}$, отвечающих наибольшим собственным значениям. Таким образом, размерность нового пространства объектов будет равна количеству кластеров K.

В общем случае матрица $\mathbf{P}$ не является ортогональной. Заметим, что для любой невырожденной матрицы $\mathbf{P}$ верно $s(\mathbf{P}) = s(\mathbf{M}\mathbf{P})$. Для учета условия ортогональности (60) найдем QR-разложение матрицы $\mathbf{P}$. Тогда ортогональная матрица $\mathbf{Q}$ является оптимальным значением $\mathbf{P}^*$. □

Оптимизация матрицы L с фиксированной матрицей P. Обозначим $\hat{\mathbf{K}} = (1/N)\mathbf{X}^{\intercal}\mathbf{P}^{\intercal}\hat{\mathbf{A}}^{-1}\mathbf{P}\mathbf{X}$. В работе [**?**] показано, что тогда задача (58) эквивалентна задаче кластеризации k-средних с заданным ядром $\hat{\mathbf{K}}$.

При фиксированной матрице $\mathbf{P}$ задача (58) принимает вид:

$$\operatorname{trace}(\mathbf{L}^{\top}\hat{\mathbf{K}}\mathbf{L}) \to \max_{\mathbf{L}\in\mathbf{R}^{m\times r}}.$$

Матрица $\hat{\mathbf{K}}$ является симметричной и неотрицательно определенной, тем самым может быть выбрана в качестве ядра.

Задача метрического обучения с динамическим выравниваем временных рядов.

Пусть объект $\mathbf{x}_i \in \mathbb{R}^n$ — временной ряд, последовательность измерений некоторой исследуемой величины в различные моменты времени. Пусть задана выборка $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ — множество объектов с известными метками классов $y_i \in \mathbb{Y}$, где $\mathbb{Y} = \{1, \ldots, K\}$ — множество меток классов.

Требуется построить точную, простую, устойчивую модель классификации $a : \mathbb{R}^n \to \mathbb{Y}$. Данную модель представим в виде суперпозиции

$$a(\mathbf{x}) = b \circ \mathbf{f} \circ G\big(\mathbf{x}, \{\mathbf{c}_e\}_{e=1}^K\big),$$

Algorithm 3 Нахождение центроида DBA($\mathbf{X}_e$, n_iter)

Вход: $\mathbf{X}_e$ — множество временных рядов, принадлежащих одному и тому же классу, n_iter — количество итераций алгоритма.

Выход: $\mathbf{c}$ — центроид множества $\mathbf{X}_e$.

1: задать начальное приближение приближение центроида $\mathbf{c}$;
2: **для** $i = 1, \ldots,$ n_iter
3: **для** $\mathbf{x} \in \mathbf{X}_e$
4: вычислить выравнивающий путь между $\mathbf{c}$ и $\mathbf{x}$
alignment($\mathbf{x}$) := DTWalignment($\mathbf{c}, \mathbf{x}$);
5: объединить поэлементно множества индексов для каждого отсчета времени
alignment := $\bigcup_{\mathbf{x}\in\mathbf{X}_e}$ alignment($\mathbf{x}$);
6: $\mathbf{c}$ = mean(alignment)

DTWalignment($\mathbf{c}$, $\mathbf{x}$)

Вход: $\mathbf{c}, \mathbf{x}$ — временные ряды.

Выход: alignment — выравнивающий путь. // *каждый индекс временного ряда* $\mathbf{x}$ *поставлен в однозначное соответствие индексу временного ряда* $\mathbf{c}$

1: построить $n \times n$-матрицу деформаций DTW
cost := DTW($\mathbf{c}, \mathbf{x}$);
2: вычислить выравнивающий путь по матрице деформаций
alignment := DTWpath(cost);

где G — процедура выравнивания временных рядов относительно центроидов классов $\{\mathbf{c}_e\}_{e=1}^K$, $\mathbf{f}$ — алгоритм метрического обучения, b — алгоритм многоклассовой классификации.

Выравнивание временных рядов. Для повышения качества и устойчивости алгоритма классификации предлагается провести выравнивание временных рядов каждого класса относительно центроида.

Пусть $\mathbf{X}_e$ — множество объектов обучающей выборки, принадлежащих одному классу $e \in \mathbb{Y}$. Центроидом множества объектов $\mathbf{X}_e = \{\mathbf{x}_i | y_i = e\}_{i=1}^m$ по расстоянию ρ назовем вектор $\mathbf{c}_e \in \mathbb{R}^n$ такой, что

$$\mathbf{c}_e = \underset{\mathbf{c}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{\mathbf{x}_i\in\mathbf{X}_e} \rho(\mathbf{x}_i,\mathbf{c}). \tag{61}$$

Для нахождения центроида предлагается в качестве расстояния между временными рядами использовать путь наименьшей стоимости [**?**, **?**], найденный методом динамической трансформации времени. Псевдокод решения оптимизационной задачи (61) приведен в алгоритме 3. Общая процедура выравнивания имеет следующий вид:

1) построить множество центроидов классов $\{\mathbf{c}_e\}_{e=1}^K$;
2) по множеству центроидов найти пути наименьшей стоимости между каждым временным рядом $\mathbf{x}_i$ и центроидом его класса $\mathbf{c}_{y_i}$;

3) по каждому пути восстановить выравненный временной ряд;
4) привести множества выравненных временных рядов к нулевому среднему и нормировать на дисперсию.

Результатом выравнивания должно стать множество выравненных временных рядов.

Метрическое обучение. Введем на множестве выравненных временных рядов расстояние Махаланобиса $d_{\mathbf{A}}$ 53. Представим матрицу трансформации $\mathbf{A}$ в виде разложения $\mathbf{A}^{-1} = \mathbf{L}^\intercal\mathbf{L}$. Матрица $\mathbf{L} \in \mathbb{R}^{p\times n}$ — матрица линейного преобразования, где p задает размерность преобразованного пространства. Если параметр $p < n$, то происходит снижение размерности признакового пространства.

Расстояние $d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)$ есть евклидово расстояние между $\mathbf{L}\mathbf{x}_i$ и $\mathbf{L}\mathbf{x}_j$:

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\intercal\mathbf{L}^\intercal\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)} = \\ = \sqrt{(\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j))^\intercal(\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j))} = \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|_2.$$

В качестве алгоритма метрического обучения в данной работе был выбран алгоритм LMNN [?]. Данный алгоритм сочетает в себе идеи метода k ближайших соседей. Первая идея заключается в минимизации расстояний между k ближайшими объектами, находящимися в одном классе. Запишем функционал качества в виде

$$Q_1(\mathbf{L}) = \sum_{j\rightsquigarrow i} \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \to \min_{\mathbf{L}},$$

где $j \rightsquigarrow i$ означает, что $\mathbf{x}_j$ является одним из k ближайших соседей для $\mathbf{x}_i$. Вторая идея состоит в максимизации расстояния между каждым объектом и его объектами-нарушителями. Объектом-нарушителем для $\mathbf{x}_i$ назовем объект $\mathbf{x}_l$ такой, что

$$\|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2 \leqslant \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 + 1, \quad \text{где } j \rightsquigarrow i. \tag{62}$$

Таким образом, необходимо минимизировать следующий функционал:

$$Q_2(\mathbf{L}) = \sum_{j\rightsquigarrow i}\sum_{l}(1 - y_{il})\big[1 + \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)\|^2 - \|\mathbf{L}(\mathbf{x}_i - \mathbf{x}_l)\|^2\big]_+ \to \min_{\mathbf{L}},$$

где $y_{il} = 1$, если $y_i = y_l$, и $y_{il} = 0$ в противном случае. Положительная срезка позволяет штрафовать только те объекты, которые удовлетворяют условию (62).

Задача метрического обучения состоит в нахождении линейного преобразования $\mathbf{f}(\mathbf{x}) = \mathbf{L}\mathbf{x}$, то есть нахождении матрицы $\mathbf{L}$ в виде решения оптимизационной задачи

$$Q(\mathbf{L}) = \mu Q_1(\mathbf{L}) + (1 - \mu)Q_2(\mathbf{L}) \to \min_{\mathbf{L}}, \tag{63}$$

где $\mu \in (0, 1)$ — весовой параметр, определяющий вклад каждого из функционалов. Задача (63) представляет собой задачу полуопределенного программирования [?] и может быть решена существующими оптимизационными пакетами.

На Рис. 7 показан принцип работы алгоритма метрического обучения LMNN по сравнению с базовым методом, использующим евклидову метрику, для случая двумерных данных. Алгоритм LMNN позволяет найти оптимальную матрицу трансформации $\mathbf{A}$, отдаляя объекты разных классов и притягивая объекты одного класса. В случае использования евклидовой метрики матрица трансформаций $\mathbf{A}$ является единичной матрицей.

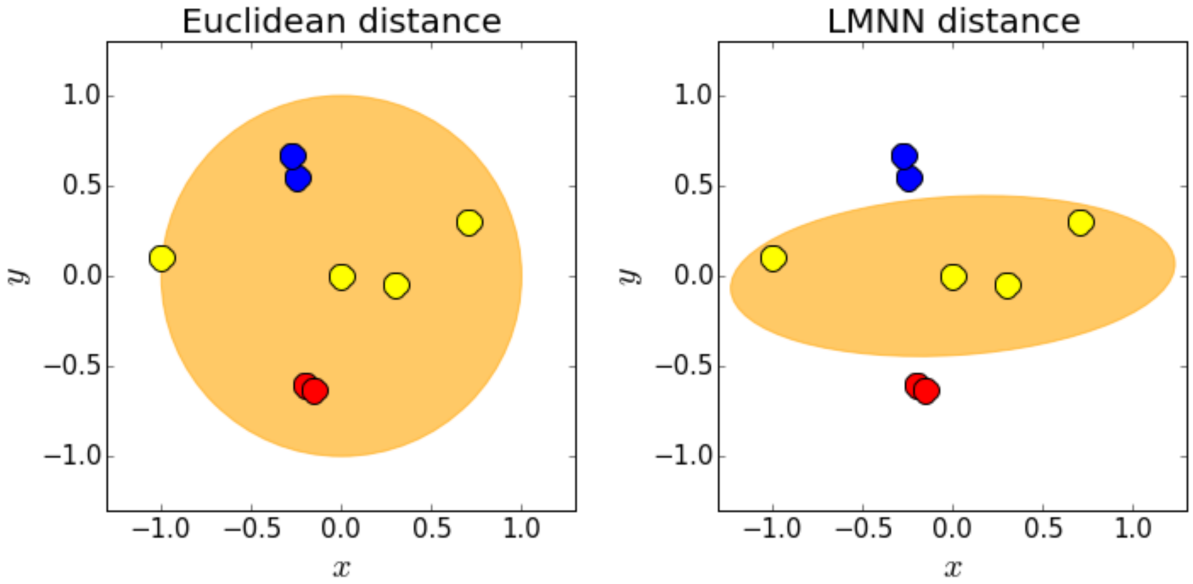

Рис. 7: Сравнение оптимальной метрики Махаланобиса алгоритма LMNN с евклидовой метрикой в двумерном случае

Классификация выравненных временных рядов в метрике Махаланобиса. Пусть $\mathbf{x} \in \mathbf{X}$ — неразмеченный временной ряд. Выравниваем временной ряд $\mathbf{x}$ относительно всех центроидов классов

$$\hat{\mathbf{x}}_e = G(\mathbf{x}, \mathbf{c}_e), \quad \text{где} \quad e \in \{1, \ldots, K\}.$$

Отнесем временной ряд к классу, для которого минимально расстояние до соответствующего центроида. В качестве расстояния используем обученную метрику Махаланобиса с фиксированной матрицей $\mathbf{A}$

$$\hat{y} = \arg\min_{e \in \mathbb{Y}} d_{\mathbf{A}}(\hat{\mathbf{x}}_e, \mathbf{c}_e).$$

После нахождения оптимальных центроидов классов и нахождения оптимальной матрицы трансформаций процедура классификации заключается в измерении расстояния между найденными центроидами и новыми неразмеченными объектами.

Для оценки качества работы алгоритма будем вычислять ошибку классификации как долю неправильно классифицированных объектов тестовой выборки $\{\mathbf{x}_i, y_i\}_{i=1}^{\hat{m}}$:

$$\text{error} = \frac{1}{\hat{m}} \sum_{i=1}^{\hat{m}} [a(\mathbf{x}_i) \neq y_i].$$

Анализ метрического пространства для задачи кластеризации.

В целях проверки работоспособности предложенного подхода проведен вычислительный эксперимент на модельных данных. Сгенерирована выборка объектов, принадлежащих одному из двух классов, в двумерном пространстве. Каждый объект принадлежит многомерному нормальному распределению. На Рис. **??** показано истинное распределение объектов, черным цветом выделены истинные центры классов и линии уровня функции распределения.

Применим к данной выборке базовый алгоритм k-средних. Результат кластеризации показан на Рис. **??**, где черным цветом выделены найденные центры классов и линии уровня функции распределения, построенной по выборочной матрице ковариаций.

Взяв за начальное приближение результаты работы алгоритма k-средних, проведем кластеризацию с помощью алгоритма адаптивного метрического обучения. Результаты работы алгоритма продемонстрированы на Рис. **??**.

На рисунках заметно улучшение результатов кластеризации. Измеренная точность кластеризации алгоритма k-средних составила 0,76, алгоритма адаптивного метрического обучения — 0,94, что говорит об эффективности данного подхода.

Таблица **??** показывает результаты вычислительного эксперимента на реальных данных. Алгоритм был применен к 5 выборкам, взятых из репозитория UCI [**?**]. Оценкой качества кластеризации служит число правильно кластеризованных объектов. При кластеризации объектов на более чем два класса возникает проблема соотнесения истинных классов с полученными кластерами. Данная проблема была формализована в виде задачи о назначениях и решена с помощью венгерского алгоритма. Вычислительный эксперимент на реальных данных показал увеличение точности кластеризации при использовании метрического обучения.

0.3 Анализ метрического пространства для задачи классификации временных рядов

Цель вычислительного эксперимента — проверить работоспособность предложенного подхода. Предполагается, что построенный алгоритм мультиклассовой классификации способен определить тип активности человека по форме сигнала акселерометра мобильного телефона.

Для проведения базового вычислительного эксперимента были подготовлены синтетические временные ряды, принадлежащие двум классам. Первый класс — синусы вида $sin(x + b)$, где параметр b определяет сдвиг каждого временного ряда. Второй класс — пилообразные функции с различными сдвигами по временной шкале. На каждый временной ряд был наложен нормальный шум. Число временных рядов каждого класса = 60. Длина каждого временного ряда $n = 50$.

Построенные центроиды классов проиллюстрированы на Рис. **??**. Из рисунка видно, что процедура корректно определяет сдвиги временных рядов.

Для того чтобы убедиться в целесообразности применения метрического обучения, данные временные ряды классифицировались в пространстве с евклидовой метрикой и в пространстве с метрикой Махаланобиса. Число ближайших соседей $k = 5$, размерность преобразованного пространства $p = 40$. Полученные ошибки классификации составили 27% для евклидовой метрики и 6% для метрики Махаланобиса.

Реальные данные [**?**] представляли собой временные ряды акселерометра мобильного телефона. Каждый из шести классов соответствовал определенной физической активности испытуемых. Для проведения вычислительного эксперимента было выбрано по 200 объектов каждого класса. Длина каждого временного ряда равнялась $n = 128$ отсчетам времени.

Построенные центроиды классов изображены на Рис. **??**. Найденные центроиды обладают периодичностью, свойственной временным рядам показаний активности человека. На Рис. **??** показаны примеры временных рядов каждого класса. Эти же временные ряды после процедуры выравнивания относительно построенных центроидов изображены на Рис. **??**.

Ошибка классификации без использования метрического обучения составила 37,5%. Алгоритм LMNN позволяет настроить параметры: число ближайших соседей k, размерность преобразованного евклидова пространства p. Для выбора оптимальных параметров воспользуемся процедурой кросс-валидации. На Рис. **??** цветом показана ошибка классификации алгоритма в зависимости от его параметров. На данной выборке алгоритм LMNN оказывается слабо чувствителен к числу ближайших соседей, и при уменьшении размерности пространства объектов ошибка классификации растет.

Настроим алгоритм LMNN со следующими параметрами: число ближайших соседей $k = 30$, размерность выходного пространства $p = 128$. Ошибка классификации составила 17,25%, что вдвое меньше ошибки классификации с использованием евклидовой метрики.

В таблице **??** представлены матрицы несоответствий результатов классификации при использовании евклидовой метрики и метрики Махаланобиса. Столбцы соответствуют истинным меткам классов объектов, строки — предсказанным меткам. Диагональное преобладание матрицы несоответствий указывает на высокую предсказательную способность алгоритма.

В таблице **??** продемонстрировано увеличение точности классификации при использовании в качестве меры расстояния метрики Махаланобиса. Пересечение i-го столбца и j-й строки отвечает изменению доли объектов класса i, отнесенных к классу j. Положительное суммарное значение диагональных элементов таблицы соответствует увеличению качества классификации. Значительное улучшение предсказания происходит при классификации первых трех классов. Данные классы соответствуют следующим видам физической активности: ходьба, ходьба вверх, ходьба вниз.

Глава 6. Порождение признаков с помощью метамоделей

Исходное пространство сигналов в задачах декодирования, а также в задачах

анализа временных рядов является крайне избыточным и неинформативным. Для извлечения информативных признаков в данной главе ставится задача порождения признакового пространства.

Постановка задачи порождения признакового пространства.

Временные ряды акселерометра образуют множество $\mathcal{S}$ сегментов $\mathbf{s}$ фиксированной длины T:

$$\mathbf{s} = [x_1, \ldots, x_T]^\intercal \in \mathbb{R}^T.$$

Необходимо построить модель классификации $f : \mathbb{R}^T \to Y$, которая будет ставить в соответствие каждому сегменту из множества $\mathcal{S}$ метку класса из конечного множества Y. Обозначим за

$$\mathcal{D} = \{(\mathbf{s}_i, y_i)\}_{i=1}^m \tag{64}$$

исходную выборку, где $\mathbf{s}_i \in \mathcal{S}$ и $y_i = f(\mathbf{s}_i) \in Y$.

Авторы предлагают построить модель f в в де суперпозиции $f = f(\mathbf{g})$. Функция $\mathbf{g} : \mathbb{R}^T \to \mathbb{R}^n$ является отображением из пространства $\mathbb{R}^T$ в признаковое пространство $\mathbb{G} \subset \mathbb{R}^n$. Имея функцию порождения признаков $\mathbf{g}$, преобразуем исходную выборку (64) в

$$\mathcal{D}_\mathbb{G} = \{(\mathbf{g}_i, y_i)\}_{i=1}^m,$$

где $\mathbf{g}_i = \mathbf{g}(\mathbf{s}_i) \in \mathbb{G}$.

Модель классификации $f = f(\mathbf{g}, \boldsymbol{\theta})$ является параметрической функцией с вектором параметров $\boldsymbol{\theta}$. Оптимальные параметры $\hat{\boldsymbol{\theta}}$ определяются оптимизацией функции ошибки классификации

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathcal{D}_\mathbb{G}, \boldsymbol{\mu}). \tag{65}$$

Вектор $\boldsymbol{\mu}$ является внешним параметром для заданной модели классификации. Примеры таких параметров и функций ошибки для различных моделей классификации приведены ниже.

Чтобы сравнить качество классификации с прошлыми результатами [**?**, **?**], в качестве метрики качества используется точность классификации:

$$\text{accuracy} = \frac{1}{m} \sum_{i=1}^{m} \left[f\left(\mathbf{g}(\mathbf{s}_i), \hat{\boldsymbol{\theta}}\right) = y_i \right]. \tag{66}$$

Модели порождения признакового пространства для временных рядов.

Цель данной работы — провести сравнение различных подходов к генерации признаков. В этом разделе проводится анализ рассматриваемых методов.

Экспертные функции. В качестве базового подхода будем использовать экспертные функции как функции порождения признаков. Экспертные функции — это некоторые статистики g_j, где $g_j : \mathbb{R}^T \to \mathbb{R}$. Признаковым описанием $\mathbf{g}(\mathbf{s})$ объекта s являются значения заданных экспертных статистик для

данного объекта

$$\mathbf{g}(\mathbf{s}) = [g_1(\mathbf{s}), \ldots, g_n(\mathbf{s})]^\intercal.$$

В работе [?] авторы предлагают использовать экспертные функции, приведенные в таблице 8. Такая процедура порождения признаков генерирует признаковое описание временного ряда $\mathbf{g}(\mathbf{s}) \in \mathbb{R}^{40}$.

Таблица 8: Примеры экспертных порождающих функций

Function description	**Formula**
Mean	$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$
Standard deviation	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2}$
Mean absolute deviation	$\frac{1}{T}\sum_{t=1}^{T}\lvert x_t - \bar{x}\rvert$
Distribution	Histogram values with 10 bins

Авторегрессионная модель. Авторегрессионная модель [?] порядка n использует параметрическую модель для аппроксимации временного ряда $\mathbf{s}$. Каждое значение временного ряда приближается линейной комбинацией предыдущих $n-1$ значений

$$x_t = w_0 + \sum_{j=1}^{n-1} w_j x_{t-j} + \varepsilon_t,$$

где ε_t — регрессионные остатки. Оптимальные параметры $\hat{\mathbf{w}}$ авторегрессионной модели используются как признаки $\mathbf{g}(\mathbf{s})$. Данные параметры минимизируют квадратичную ошибку аппроксимации временного ряда и предсказания модели

$$\mathbf{g}(\mathbf{s}) = \hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \left(\sum_{t=n}^{T} \|x_t - \hat{x}_t\|^2 \right). \tag{67}$$

Задача (67) эквивалентна задаче линейной регрессии. Поэтому для каждого временного ряда s необходимо решить задачу линейной регрессии размера n. Пример аппроксимации временного ряда авторегрессионной моделью представлен на Рис. **??**.

Анализ сингулярного спектра. Альтернативной гипотезой порождения признакового пространства для временного ряда является анализ сингулярного спектра (Singular Spectrum Analysis, SSA) [?]. Для каждого временного ряда $\mathbf{s}$ из исходной выборки $\mathcal{D}$ строится траекторная матрица:

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ x_2 & x_3 & \ldots & x_{n+1} \\ \ldots & \ldots & \ldots & \ldots \\ x_{T-n+1} & x_{T-n+2} & \ldots & x_T \end{pmatrix}.$$

Здесь ширина окна n является внешним структурным параметром. Сингулярное разложение матрицы $\mathbf{X}^\intercal\mathbf{X}$:

$$\mathbf{X}^\intercal\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal,$$

где $\mathbf{U}$ — унитарная матрица и $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ причём λ_i собственные значения $\mathbf{X}^\intercal\mathbf{X}$. Признаковое описание объекта $\mathbf{s}$ задаётся спектром матрицы $\mathbf{X}^\intercal\mathbf{X}$:

$$\mathbf{g}(\mathbf{s}) = [\lambda_1, \ldots, \lambda_n]^\intercal.$$

Spline Approximation. Предлагаемый метод аппроксимирует временные ряды с помощью сплайнов [**?**]. Сплайн определяется его параметрами: узлами и коэффициентами. Предполагается, что узлы сплайна $\{\xi_\ell\}_{\ell=0}^M$ равномерно распределены по временной оси. Кусочные модели, построенные на отрезках $[\xi_{\ell-1}; \xi_\ell]$, заданы коэффициентами $\{\mathbf{w}_\ell\}_{\ell=1}^M$. Оптимальные параметры сплайна являются решением системы с дополнительными условиями равенства производных до второго порядка включительно на концах отрезков. Обозначим каждый отрезок-сегмент $p_i(t)$ $i = 1, \ldots, M$ и весь сплайн $S(t)$. Тогда система уравнений принимает вид

$$S(t) = \begin{cases} p_1(t) = w_{10} + w_{11}t + w_{12}t^2 + w_{13}t^3, & t \in [\xi_0, \xi_1], \\ p_2(t) = w_{20} + w_{21}t + w_{22}t^2 + w_{23}t^3, & t \in [\xi_1, \xi_2], \\ \ldots & \ldots \\ p_M(t) = w_{L0} + w_{M1}t + w_{M2}t^2 + w_{M3}t^3, & t \in [\xi_{M-1}, \xi_M], \end{cases}$$

$$\begin{aligned} S(\xi_t) &= x_t, \quad t = 0, \ldots, M, \\ p_i'(\xi_i) &= p_{i+1}'(\xi_i), p_i''(\xi_i) = p_{i+1}''(\xi_i), \quad i = 1, \ldots, M-1, \\ p_i(\xi_{i-1}) &= x_{i-1}, p_i(\xi_i) = x_i, \quad i = 1, \ldots, M. \end{aligned}$$

Объединение всех параметров сплайна задаёт признаковое описание временного ряда:

$$\mathbf{g}(\mathbf{s}) = [\mathbf{w}_1, \ldots, \mathbf{w}_M]^\intercal.$$

Рис. **??** показывает аппроксимацию временного ряда с использованием модели сплайнов. По сравнению с авторегрессионной моделью сплайны строят более гладкую аппроксимацию, используя такое же количество параметров.

0.4 Классификация временных рядов в порожденном признаковом пространства

Для классификации временных рядов будем использовать подход один против всех. Для каждого класса обучается бинарный классификатор, и на стадии предсказания объект классифицируется согласно наиболее уверенному классификатору. Использовались три модели классификации: логистическая регрессия, SVM и случайный лес.

Логистическая регрессия. Оптимальные параметры модели $\hat{\mathbf{w}}, \hat{b}$ в случае логистической регрессии определяются минимизацией функции ошибки (65)

$$L(\boldsymbol{\theta}, \mathcal{D}_{\mathbb{G}}, \mu) = \sum_{i=1}^{m} \log\big(1 + \exp(-y_i[\mathbf{w}^{\intercal}\mathbf{g}_i + b])\big) + \frac{\mu}{2}\|\mathbf{w}\|^2, \text{ where } \boldsymbol{\theta} = \begin{bmatrix}\mathbf{w}\\ b\end{bmatrix}.$$

Решающее правило $f(\mathbf{g}, \boldsymbol{\theta})$ — знак линейной комбинации описания объекта $\mathbf{g}$ и параметров $\hat{\boldsymbol{\theta}}$

$$\hat{y} = f(\mathbf{g}, \hat{\boldsymbol{\theta}}) = \operatorname{sgn}(\mathbf{g}^{\intercal}\hat{\mathbf{w}} + \hat{b}).$$

SVM. Оптимизационная задача метода SVM имеет вид

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix}\hat{\mathbf{w}}\\ \hat{b}\\ \hat{\xi}\end{pmatrix} = \arg\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \mu\sum_{i=1}^{m}\xi_i, \text{ s.t. } y_i\left(\mathbf{w}^{\intercal}\mathbf{g}_i + b\right) \geqslant 1 - \xi_i,$$

$$\xi_i \geqslant 0, \quad 1 \leqslant i \leqslant m.$$

Целевая функция соответствует функции ошибки классификации $L(\boldsymbol{\theta}, \mathcal{D}_{\mathbb{G}}, \mu)$. Предсказание для нового объекта вычисляется аналогично $\hat{y} = \operatorname{sgn}(\mathbf{g}^{\intercal}\hat{\mathbf{w}} + \hat{b})$.

Случайный лес. Случайный лес использует идею бэггинга. Идея состоит в построении многих слабых, неустойчивых классификаторов на подвыборках с возвращениями и усреднения их предсказаний. Метод предполагает использование в качестве базовых классификаторов моделей с низким смещением и высокой дисперсией. Усреднение позволяет уменьшить дисперсию. В случае случайного леса базовой моделью выступают решающие деревья. Идея бэггинга используется не только для самих объектов, но и для множества признаков. В данном случае предсказание для нового объекта получается усреднением всех предсказаний отдельных деревьев:

$$\hat{y} = \operatorname{sgn}\left(\frac{1}{B}\sum_{i=1}^{B}\operatorname{pred}(\mathbf{g}_i)\right),$$

где B — количество деревьев в композиции.

Анализ порожденных признаковых пространств.

В данной работе эксперименты проводились на двух наборах данных временных рядов акселерометра мобильного телефона: WISDM [**?**] и USC-HAD [**?**]. Акселерометр мобильного телефона проводит измерение ускорения по трём осям с частотой 100 Hz. Данные WISDM содержат 4321 временной ряд. Каждый временной ряд принадлежит к одному из 6 классов. Данные USC-HAD содержат 13620 временных рядов, принадлежащих одному из 12 классов. В таблице **??** представлено распределение временных рядов по классам для каждого датасета.

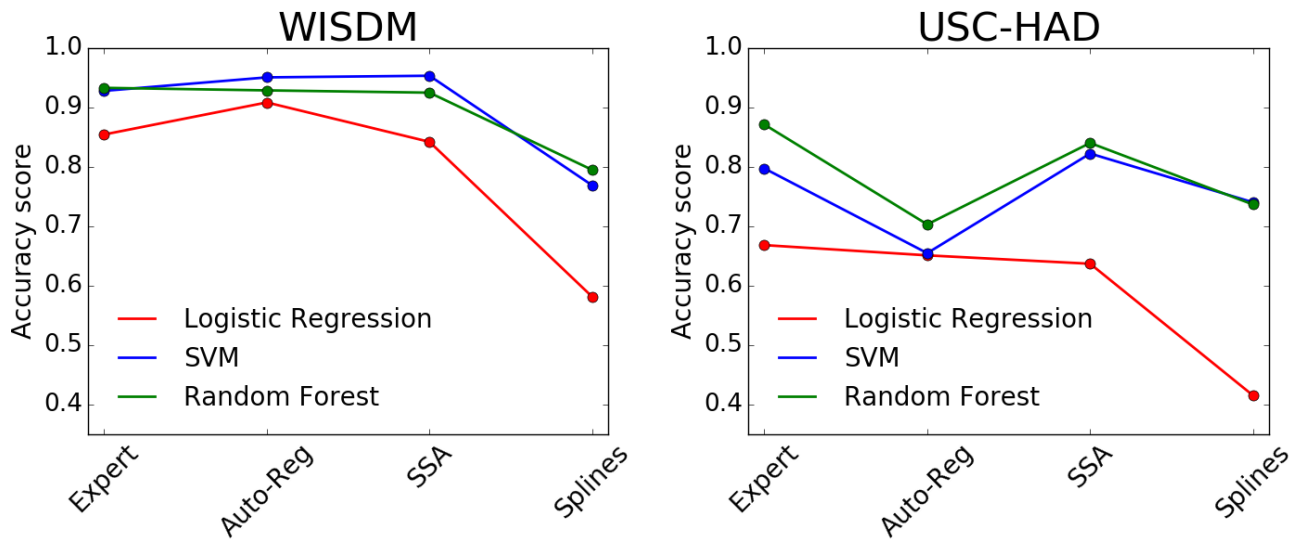

Рис. 8: Мультиклассовая точность классификации для различных порожденных признаковых пространств

Длина временного ряда равна 200. На Рис. **??** представлен пример одного из временных рядов.

В эксперименте для каждого набора данных были порождены признаки одним из методов: экспертные функции, авторегрессионная модель, SSA и сплайны. Для каждой процедуры порождения признакового описания настраивались три модели классификации: логистическая регрессия, SVM и случайный лес. Внешние структурные параметры (длина авторегрессионной модели n, ширина окна SSA n, число узлов сплайна M) настраивались процедурой кросс-валидации:

$$CV(K) = \frac{1}{K}\sum_{k=1}^{K} L(f_k, \mathcal{D} \setminus \mathcal{C}_k),$$

где $\mathcal{C}_k$ —$\frac{K-1}{K}$ доля от всей выборки, используемая для обучения модели f_k. Гиперпараметры $\boldsymbol{\mu}$ моделей классификации были настроены той же процедурой кросс-валидации.

Первый подход к порождению признаков временных рядов — экспертные функции. Основной недостаток такого подхода необходимость экспертного задания функций и возможности их вычисления для конкретного набора данных.

Авторегрессионная модель требует задания параметра длины модели n. Процедура кросс-валидации дала наибольшее качество при $n = 20$ для обоих наборов данных.

Модель SSA была настроена аналогичной процедурой выбора оптимальных гиперпараметров. Конечная модель имела ширину окна $n = 20$.

Для аппроксимации временных рядов кубическими сплайнами [**?**] использовалась библиотека *scipy*. Узлы сплайнов $\{\xi_\ell\}_{\ell=1}^{M}$ были распределены равномерно по временной оси. Значение параметра M было подобрано на кросс-валидации.

Для обоих наборов данных процедуры порождения признаковых описаний дали следующие количества признаков: экспертные функции — 40; авторегрессионная модель — 60; анализ сингулярного спектра — 60; сплайны — 33.

Таблица 9: Бинарная точность классификации для данных WISDM с использованием рассматриваемых алгоритмов: EX — Expert, AR — Auto-Reg, SSA and SPL for Splines

	Logistic Regression				**Random Forest**				**SVM**			
	EX	AR	SSA	SPL	EX	AR	SSA	SPL	EX	AR	SSA	SPL
All	0.85	0.91	0.84	0.58	0.93	0.93	0.92	0.79	0.93	0.95	0.95	0.77
Standing	0.99	0.98	1.00	0.95	1.00	0.99	1.00	0.99	0.99	0.98	1.00	0.96
Walking	0.91	0.96	0.86	0.61	0.96	0.97	0.95	0.86	0.96	0.98	0.98	0.84
Upstairs	0.91	0.95	0.91	0.89	0.96	0.96	0.96	0.90	0.96	0.98	0.97	0.89
Sitting	0.99	0.98	1.00	0.99	1.00	0.99	1.00	1.00	0.99	0.98	1.00	1.00
Jogging	0.98	0.99	0.99	0.80	0.99	0.99	0.99	0.92	0.99	0.99	0.99	0.93
Downstairs	0.93	0.96	0.94	0.92	0.96	0.97	0.96	0.92	0.96	0.98	0.97	0.92

Таблица 10: Бинарная точность классификации для данных USC-HAD с использованием рассматриваемых алгоритмов: EX — Expert, AR — Auto-Reg, SSA and SPL for Splines

	Logistic Regression				**Random Forest**				**SVM**			
	EX	AR	SSA	SPL	EX	AR	SSA	SPL	EX	AR	SSA	SPL
All	0.67	0.65	0.64	0.41	0.87	0.70	0.84	0.74	0.80	0.65	0.82	0.74
Standing	0.94	0.94	0.92	0.89	0.98	0.94	0.97	0.98	0.95	0.94	0.97	0.96
Elevator-up	0.94	0.94	0.93	0.92	0.95	0.95	0.95	0.95	0.93	0.94	0.94	0.93
Walking-forward	0.87	0.87	0.89	0.70	0.97	0.89	0.96	0.88	0.95	0.87	0.97	0.91
Sitting	0.98	0.95	0.94	0.96	0.99	0.96	0.98	0.99	0.98	0.96	0.99	0.99
Walking-downstairs	0.95	0.93	0.93	0.90	0.99	0.96	0.98	0.95	0.98	0.93	0.98	0.96
Sleeping	1.00	0.98	0.99	1.00	1.00	0.98	1.00	1.00	1.00	0.98	1.00	1.00
Elevator-down	0.94	0.94	0.94	0.91	0.95	0.95	0.95	0.95	0.93	0.94	0.94	0.93
Walking-upstairs	0.94	0.95	0.93	0.92	0.98	0.95	0.98	0.96	0.98	0.95	0.98	0.96
Jumping	0.99	0.99	1.00	0.97	1.00	0.99	1.00	0.99	1.00	0.99	0.97	0.99
Walking-right	0.91	0.90	0.91	0.86	0.97	0.92	0.96	0.92	0.96	0.90	0.97	0.93
Walking-left	0.89	0.91	0.90	0.88	0.97	0.93	0.97	0.93	0.95	0.91	0.97	0.93
Running	0.99	0.99	0.99	0.92	1.00	0.99	1.00	0.97	1.00	1.00	0.95	0.98

На Рис. 8 показано качество классификации (66) для двух наборов данных. Для данных WISDM сплайны дали самое слабое качество классификации. Результаты для экспертных функций, авторегрессионной модели и SSA схожи. Для данных USC-HAD результат более восприимчив к выбору модели классификации. Для обоих наборов данных логистическая регрессия продемонстрировала наименьшее качество, SVM и случайный лес показали почти одинаковое качество. Для набора данных USC-HAD модель с использованием аппроксимации сплайнами показала сравнимое с другими методами качество.

В таблицах 9 и 10 представлены результаты классификации (66) для каждого класса в отдельности. Первая строка в обеих таблицах демонстрирует точность по всем классам для каждой модели и процедуры генерации признаков. Следующие строки соответствуют бинарным точностям по каждому из классов.

Для данных WISDM лучшее качество имеют наименее активные классы, такие как Standing и Sitting. Для USC-HAD заметного выделения качества для определенных классов не наблюдается.

Также был проведён эксперимент с использованием объединённого множества всех 193 сгенерированных признаков. Результаты представлены на Рис. **??**. Соответствие между номера классов и видами активности приведено в таблице **??**. Объединение признаков для обучения одной модели позволило увеличить качество. Для данных WISDM все точности классификации по классам больше 97%, а для USC-HAD выше 93%.

ПРО ЗАКЛЮЧЕНИЕ

Публикации соискателя по теме диссертации

Публикации в журналах из списка ВАК.

1. Исаченко Р. В., Стрижов В. В. Метрическое обучение в задачах мультиклассовой классификации временных рядов // Информатика и её применения, 2016. Т. 10. № 2. С. 48–57.
2. Isachenko R. et al. Feature Generation for Physical Activity Classification // Artificial Intelligence and Decision Making, 2018. № 3. С. 20–27.
3. Isachenko R. V., Strijov V. V. Quadratic programming optimization with feature selection for nonlinear models // Lobachevskii Journal of Mathematics, 2018. Т. 39. № 9. С. 1179–1187.
4. Isachenko R. V., Vladimirova M. R., Strijov V. V. Dimensionality Reduction for Time Series Decoding and Forecasting Problems //DEStech Transactions on Computer Science and Engineering, 2018. №. optim.
5. Исаченко Р.В., Яушев Ф.Ю., Стрижов В.В. Модели согласования скрытого пространства в задаче прогнозирования // Системы и средства информатики, 2021. Т. 31. № 1.

Прочие публикации.

6. Исаченко Р. В., Катруца А. М. Метрическое обучение и снижение размерности пространства в задачах кластеризации // Машинное обучение и анализ данных, 2016. Т. 2. № 1. С. 17–25.